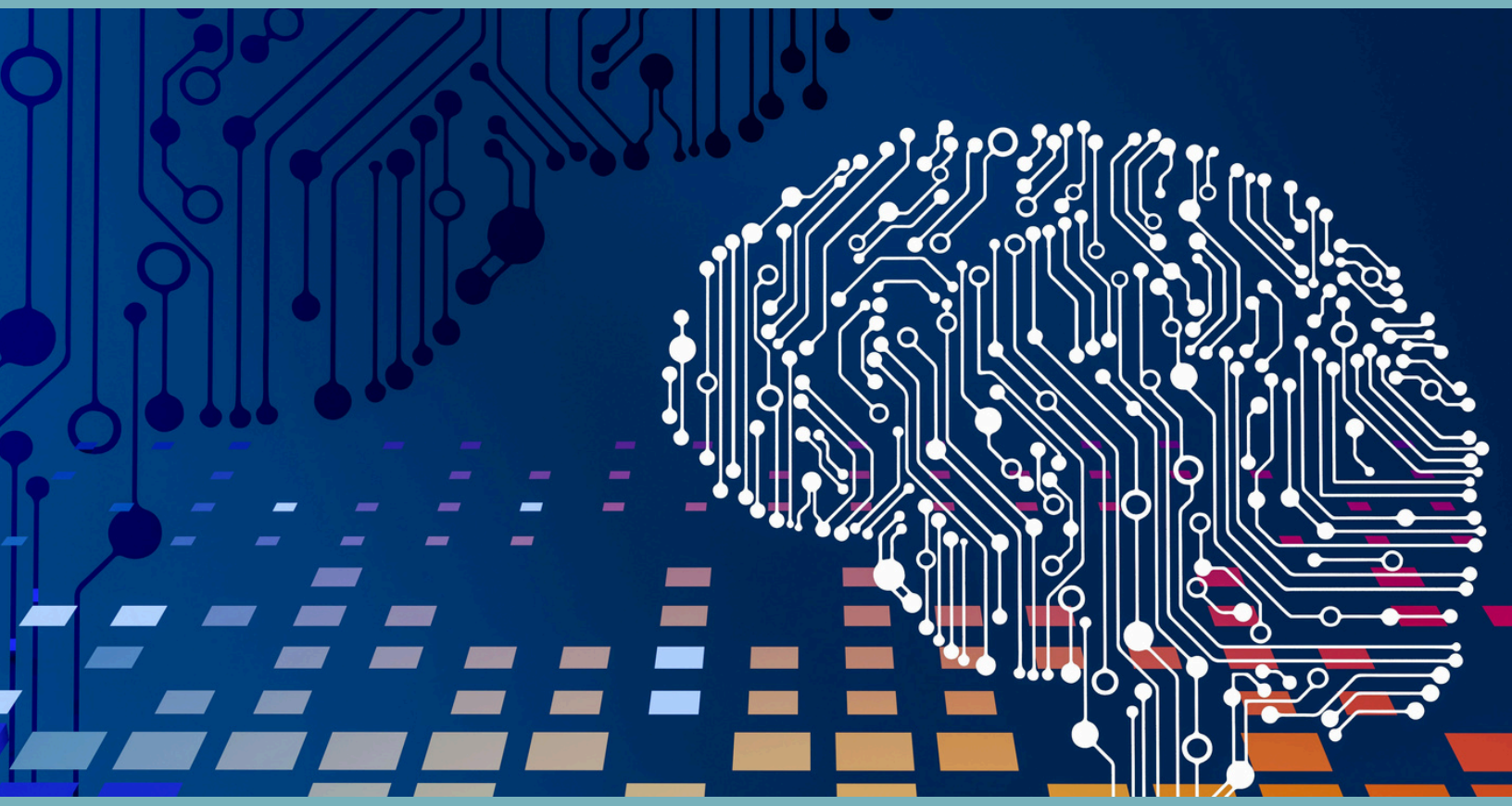
REPORT ON

ARTIFICIAL INTELLIGENCE AND CIVIL LIABILITY



BRITISH COLUMBIA LAW INSTITUTE

APRIL 2024

Report on Artificial Intelligence and Civil Liability

A Report Prepared for the British Columbia Law Institute by the Members of the Artificial Intelligence and Civil Liability Project Committee

BCLI Report no. 96 **April 2024**

Disclaimer

The information and commentary in this publication is not offered as legal advice. It refers only to the law at the time of publication, and the law may have since changed. BCLI does not undertake to continually update or revise each of its publications to reflect post-publication changes in the law.

The British Columbia Law Institute and its division, the Canadian Centre for Elder Law, disclaim any and all responsibility for damage or loss of any nature whatsoever that any person or entity may incur as a result of relying upon information or commentary in this publication.

You should not rely on information in this publication in dealing with an actual legal problem that affects you or anyone else. Instead, you should obtain advice from a qualified legal professional concerning the particular circumstances of your situation.

Cover photograph by Steve Johnson on Unsplash. Cover design by Ken Chau.

Published in Vancouver on unceded Coast Salish homelands, including the territories of the xʷməθkwəy̓əm (Musqueam), Skwxwú7mesh (Squamish), and Səl̓ílwətaʔ/ Selilwitulh (Tsleil-Waututh) Nations.

British Columbia Law Institute

1822 East Mall, University of British Columbia, Vancouver, BC, Canada V6T 1Z1

Voice: (604) 822-0142 Fax: (604) 822-0144 E-mail: bcli@bcli.org
WWW: https://www.bcli.org

The British Columbia Law Institute was created in 1997 by incorporation under the provincial *Society Act*. Its purposes are to:

- promote the clarification and simplification of the law and its adaptation to modern social needs,
- promote improvement of the administration of justice and respect for the rule of law, and
- promote and carry out scholarly legal research.

This project was made possible with the sustaining financial support of the Law Foundation of British Columbia and the Ministry of Attorney General for British Columbia. The Institute gratefully acknowledges the support of the Law Foundation and the Ministry for its work.

Introductory Note

Report on Artificial Intelligence and Civil Liability

A refrain constantly heard today is that artificial intelligence is transforming the world as we know it. While opinions may differ on whether the "transformation" paradigm is an accurate depiction or an exaggeration, there is unquestionably a need in most fields of human activity and human institutions to adapt to increasing levels of automated decision-making. As a human institution, law is not immune. Adaptation requires re-assessment of fundamental legal premises based on human perceptions and experience to determine whether these premises remain valid and, if so, how they may be applied in an environment where increasingly capable autonomous machines bring benefits but also new sources of risk.

The *Report on Artificial Intelligence and Civil Liability* addresses the potential for artificial intelligence to cause harm to persons, property, and other interests protected by the current law of torts and offers answers to the questions "Who is, or should be, liable for choices made by intelligent machines operating autonomously, and when?"

The report begins with an acknowledgement that there is no single, comprehensive definition of artificial intelligence and draws on a cross-section of existing definitions to provide a non-technical explanation of what that term is generally understood to comprise. It analyzes why difficulties emerge in attempting to apply legal rules that developed over centuries to provide redress for wrongful conduct by human tortfeasors to situations where harm results from the autonomous operation of an artificial "mind." It sets out recommendations for adapting existing tort principles to address these novel circumstances, which include the potential for serious harm to individuals and classes of persons through "algorithmic discrimination," the replication and amplification of undetected biases hidden in algorithms and data.

As its recommendations are designed to be implemented either through legislation or judicial decision, this report can serve to aid both legislatures and civil courts, regardless of which is the earlier to grapple with the issues they cover and establish guideposts for civil justice in the age of artificial intelligence.

Edward L. Wilson
Chair,
British Columbia Law Institute
April 2024

Artificial Intelligence and Civil Liability Project Committee

The interdisciplinary Artificial Intelligence and Civil Liability Project Committee was formed in November 2021 to assist BCLI in developing recommendations on how the rules of tort law should adapt to respond to harm caused by autonomous artificial intelligence systems.

The members of the committee are:

Prof. Robert G. Howell - Chair
Faculty of Law
University of Victoria
BCLI Member Emeritus

Dylan Merrick
Former Member,
BCLI Board of Directors

Dr. Cindy Grimm
School of Mechanical, Industrial, and Manufacturing Engineering
Oregon State University

Dr. Kristen Thomasen
Peter A. Allard School of Law
University of British Columbia

Cynthia Khoo
Barrister and Solicitor
Former Senior Associate, Center on Privacy & Technology at Georgetown Law
Research Fellow, Citizen Lab, University of Toronto

Darin Thompson
Liaison Observer on behalf of Ministry of Attorney General
Chief Policy Officer
Civil and Criminal Policy Division
Justice Services Branch
Ministry of Attorney General

Maya Medeiros
Barrister and Solicitor
Norton Rose Fulbright LLP

Dr. Teresa S.M. Tsang
Director, VGH-UBC Echocardiography Lab and AI Echo Core Lab
Executive Director, Vancouver Coastal Health Research Institute
Associate Head Research, Department of Medicine and Professor, Cardiology
Associated Dean, Research,
Faculty of Medicine,
University of British Columbia

Gregory G. Blue, K.C. (senior staff lawyer, BCLI) is the project manager.

For more information, visit us on the World Wide Web at
https://www.bcli.org/project/artificial-intelligence-and-civil-liability-project/

TABLE OF CONTENTS

Acknowledgements

The British Columbia Law Institute wishes to thank the members of the Artificial Intelligence and Civil Liability Project Committee for their dedication and generous contribution of time and expertise throughout the two years of intense deliberations culminating in this report. Without them, the project could not have been carried out. A special debt is owed to Prof. R.G. Howell for chairing the Project Committee. We are also grateful for the participation of Darin Thompson, Chief Policy Officer in the Civil and Criminal Policy Division of the Ministry of Attorney General, who observed the Project Committee in action and was the liaison with the Ministry throughout the project.

We thank all the individuals and organizations that responded to the consultation paper issued in the course of the project, including all contributors to the collegial response provided on behalf of the B.C. Branch of the Canadian Bar Association. The extensive and detailed submissions made in response to the consultation paper greatly assisted the Project Committee and are sincerely appreciated.

We express our special gratitude to Dr. Kevin Layton-Brown of the UBC ICICS Centre for Artificial Intelligence Decision-making and Action (CAIDA), Dr. Randy Goebel of the Alberta Machine Intelligence Institute, Prof. Gillian K. Hadfield of the Schwartz Reisman Institute for Technology and Society, and Alexis Leblanc-Roy of the Cyberjustice Laboratory, University of Montréal for their assistance in disseminating the consultation paper within the AI research and industry community and encouraging its members to engage with us regarding it.

The Artificial Intelligence and Civil Liability Project was generously funded by the Ministry of Attorney General, and would not have been possible otherwise. We remain grateful to the Ministry for its interest in and support of our work.

The Institute acknowledges the contribution of members of the Institute's staff to the project and the preparation of this report: Karen Campbell, Greg Blue, K.C., Alison Wilkinson, Megan Vis-Dunbar, Shauna Nicholson, Ken Chau. The contribution of Dan Lee, a law student assisting in the course of the project, is also acknowledged.

Executive Summary

Artificial intelligence is pervasive in today's world. Decisions, predictions and recommendations made by artificial intelligence affect individual lives to an ever-increasing extent.

The benefits of artificial intelligence are great, but they also come with risks. When the risks materialize, harm to persons and property may result. Avoidable harm to persons and property brings the law of tort into play.

Tort is the branch of law concerned with non-contractual civil wrongs. Its principles evolved in order to deal with harmful human conduct. Involvement of artificial intelligence as a causal factor leading to harm complicates the application of those principles to determine who is legally responsible, when and for what. Many questions concerning rights and liabilities in that context have yet to be answered because these situations are new, and the law relating to artificial intelligence is unsettled.

The Artificial Intelligence and Civil Liability Project

BCLI initiated the Artificial Intelligence and Civil Liability Project in late 2021. The objectives of the project were:

- to identify how the common law rules of tort need to be adapted to provide just and adequate civil remedies for harm to persons, property, and other legal interests resulting from the operation of autonomously functioning artificial intelligence; and

- to develop and publish law reform recommendations addressing that context.

BCLI was assisted in carrying out the project by the Artificial Intelligence and Civil Liability Project Committee, an interdisciplinary group reflecting expertise in computer science, engineering, and medicine as well as law.

A consultation paper was issued in July 2023 to gather the views of stakeholders and the general public on tentative law reform recommendations developed by the Project Committee. The responses to the consultation paper greatly assisted the process of reaching the final recommendations set out in this report.

Outline of the Report

Chapter 1

Chapter 1 provides general background on the Artificial Intelligence and Civil Liability Project and the reasons why BCLI undertook it.

Chapter 2

Chapter 2 reviews various definitions of artificial intelligence explains and describes its various forms, including machine learning. As there is no single, universally accepted definition, the Project Committee developed a "working description" of artificial intelligence for the purposes of the project, building upon elements common to many definitions and its generally recognized forms.

Chapter 2 then describes characteristics and phenomena associated with artificial intelligence that have bearing on the legal issues discussed in the later chapters. One of these is "autonomy," a characteristic that artificial intelligence systems commonly possess to varying extents. It is explained as the ability of a system or device to interact with its environment and pursue objectives assigned to it without continuous human input or control.

Limited explainability is another characteristic of some artificial intelligence systems, especially those that rely on machine learning and artificial neural networks. Systems capable of machine learning are not fully dependent on programmed instructions by humans, but can base outputs on inferences from patterns they detect in data and improve their performance by trial and error. They can re-use the inferences from data to which they have been exposed to process and analyze new data. The abstract models the systems create internally to represent the relationships they find in the data they process may not be interpretable by humans. Thus, the process from input to output on these systems that are heavily dependent on data is not always explainable. This is why the expression "black box" is often applied to artificial intelligence in the media and popular literature.

While "black box" is a misleading and unhelpful term, the decision-making of data-dependent systems cannot always be validated on the basis of a step-by-step chain of reasoning. If the way a system produces its outputs cannot be explained, the way it will respond to a specific input is less predictable. Systems functioning on the basis of machine learning have been called "unpredictable by design."

Artificial intelligence systems have sometimes displayed highly unpredicted, original behaviour in pursuit of their objectives that is referred to as "emergence." Emergence has a double aspect. It is beneficial much of the time, solving seemingly intractable problems and astounding even the designers and programmers of the systems that display it. On occasion, however, the same capabilities that allow artificial intelligence to generate unanticipated but highly desirable results may also generate unintended, harmful results.

When the outputs of artificial intelligence result in harm to persons, property, or other legally protected interests, implications of civil liability will arise under tort law. The rest of the report explores those implications.

Chapter 3

Chapter 3 deals with the legal basis for imposing civil liability when harm results from the operation of artificial intelligence. The point is made at the beginning that legal responsibility for the performance of an artificial intelligence system must rest ultimately with the humans or corporate entities behind the system, because systems themselves do not have the means of compensating their victims. The chapter goes on to cover basic principles of tort law that relate to the rest of the report, such as fault, intention, and the elements of negligence.

Various approaches to civil liability for harm from artificial intelligence that have been proposed in the common law world are then reviewed and evaluated along with ones proposed within the EU, which is ahead of much of the world in addressing legal frameworks for use of artificial intelligence. These include strict liability, product liability, notional agency, vicarious liability, legal neutrality (applying the same legal standards of reasonable conduct to harm-causing artificial intelligence as are applied to human tortfeasors), and a sliding scale of liability regimes depending on levels of risk.

The report recommends retention of a fault-based regime for addressing harm caused by artificial intelligence rather than introducing one of strict liability. Under strict liability, a plaintiff needs only to prove damage and causation. Absence of fault on the part of the defendant is irrelevant. Among the several grounds presented for rejecting strict liability is that it reduces the incentive for continual improvement of standards in the design, development, and use of artificial intelligence, because defendants would be held liable regardless of the degree of care they exercise. Apart from cases of harm caused intentionally by means of artificial intelligence, a torts regime under which civil liability arises only when damage has resulted from a failure to meet a reasonable standard of care is more likely to achieve a fair balance between risk and the benefits of technological innovation.

A division of potential defendants into two classes is proposed: those "upstream" in the chain of events leading to litigation and those situated "downstream." Chapter 3 contains separate recommendations on the legal basis for the liability of upstream and downstream defendants, respectively.

Upstream defendants are those involved in the design, development, training and testing of artificial intelligence systems prior to the point at which the systems are placed on the market or deployed in an actual use case. Downstream defendants are those who use an artificial intelligence system for their own purposes or have control over the risks of its operation once the system has been deployed or released for actual, real-world use following the design, development, training, and testing phases.

The two classes are not mutually exclusive and could overlap. For example, a company that develops an artificial intelligence system for its own use would be both an upstream and downstream defendant.

The recommendation on the liability of *upstream* defendants for unintended harm is that it should be based on an adaptation of product liability principles. In Canadian tort law, product liability is a branch or offshoot of negligence with well-established principles surrounding the scope of the duty of care of manufacturers and others in the supply chain. The duty of care of a manufacturer extends to anyone who may come into contact with a product after it enters the stream of commerce and thus may be affected by its intrinsic defects or hazards.

Developers of a complete, integrated artificial intelligence system may be compared to manufacturers of a complex product and should owe a duty of care accordingly. Designers and developers of components, including an artificial intelligence module that is incorporated into an integrated software system or device, should owe a duty of care analogous to that owed by suppliers of components of a complex product, who are considered manufacturers in their own right for the purposes of product liability.

The corresponding recommendation concerning *downstream* defendants is that their liability for unintended harm should be governed essentially by ordinary negligence principles.

The later chapters contain additional recommendations modifying the application of negligence principles to both upstream and downstream defendants in order to address issues that will arise in tort claims for damage stemming from the operation of artificial intelligence.

Chapter 4

Chapter 4 addresses the formidable obstacles that may face a plaintiff in proving causation and fault if the damage is linked with non-human, autonomous decision-making by artificial intelligence. There will be many potential defendants, because a complete artificial intelligence system typically integrates numerous elements from multiple sources.

Outward opacity of many systems, together with the tendency of the software industry to treat algorithms and other aspects of the design of systems as proprietary secrets, has emerged as a problem in civil litigation. Defendants other than the designers of the underlying algorithms and handlers of training data may be as much in the dark as the plaintiff regarding information about the system that is crucial to a fair adjudication of a damage claim.

Even the designers of an artificial intelligence system may be unable to explain how systems that are highly data-dependent and involve artificial neural networks reach particular outputs. Given this, pre-trial discovery processes alone may be inadequate to redress the "informational asymmetry" between plaintiffs and defendants. If the route from input to output is not fully explainable even by the designers and programmers of a system, plaintiffs have little hope of being able to put forward a coherent, provable theory of causation linking a breach of legal duty by the defendant to the harm incurred.

Res ipsa loquitur might formerly have had a role to play in such a situation in Canadian common law jurisdictions, but the Supreme Court of Canada has held that *res ipsa loquitur* is obsolete and may no longer be invoked as an evidentiary mechanism in aid of proof of negligence. Numerous academic writers in the common world have reached the conclusion that these informational obstacles in the way of proof of causation and fault warrant some mechanism to maintain balance in the litigation process in cases involving artificial intelligence, as have EU policymakers. The problem has also concerned policymakers in the EU. The European Commission's *AI Liability Directive Proposal* is chiefly concerned with creating a more equal playing field for claimants (plaintiffs) and defendants in fault-based claims under the national law of EU Member States.

Our report recommends that in tort claims for harm resulting from artificial intelligence in operation, courts should be justified in drawing an inference of a causal link between a failure to exercise reasonable care in the design, development, training, testing, or use of an artificial intelligence system and the harm incurred by the plaintiff if three conditions are satisfied:

- Harm is proven to have been caused by the output of an artificial intelligence system functioning alone or as a component of an integrated system;

- The evidence as a whole does not show either that reasonable care was exercised in the design, development, training, testing, or use of the system, or an explanation for the behaviour of the system in the circumstances that is consistent with exercise of reasonable care; and

- The characteristics of the system (e.g., opacity) make it unreasonable to expect the plaintiff to identify a specific act or omission on the part of a specific defendant that caused or contributed to causing the system to occasion the harm.

The inference of causation could be rebutted with respect to a particular defendant by evidence showing that the defendant in question exercised reasonable care to prevent the system from causing harm of the nature incurred by the plaintiff.

Foreseeability of harm would be a prerequisite to liability in negligence-based claims arising from harm caused by artificial intelligence. Reasonable foreseeability as conventionally understood and applied by Canadian courts can become unstuck, however, when the claim is based on non-human decision-making.

Described in case law in terms of what a reasonable person would see as a "real risk" that is not "far-fetched," or in terms of a "natural result" of an act or omission, reasonable foreseeability is inextricably linked to human experience and human perceptions of cause and effect. The reactions of a non-human decision-making process to an infinite range of potential inputs are outside that experience. Under the conventional tests, harmful emergence will seldom be found to have been reasonably foreseeable. This would result in victims incurring harm through artificial intelligence having less protection under tort law than victims of human tortfeasors.

To avoid this result, the report recommends that foreseeability of risk in relation to the use of artificial intelligence should not be thought of in terms of particular outcomes considered in isolation, but in terms of unpredictability and emergence being known risks that potentially give rise to unknown ones.

The report also recommends that in making determinations regarding what was reasonably foreseeable as a risk that might materialize in an emergent and unpredictable manner, courts should take into account the known attributes of the system in question at the relevant time, the use cases for which the system was intended, and known or predictable alternate use cases (including predictable misuse). These are considerations that delineate the scope of potential harm that designers,

developers, and operators should be expected to consider in assessing risk and taking measures to avert it.

Chapter 5

Chapter 5 is concerned with how the standard of care should be set in claims based on harm arising from the operation of artificial intelligence. In negligence cases, courts can look to statutory and regulatory requirements affecting the activity in issue in a negligence case and to evidence of the accepted or customary standard in the industry or profession in question. These are relevant considerations in setting the standard of care in an individual case, but not decisive regarding it. Breach of a statute or regulation applicable to the defendant or the defendant's activities in issue does not automatically lead to a finding of liability, nor will compliance necessarily avoid it. The same is true with respect to accepted or standard practice within an industry.

There is little regulation of artificial intelligence in Canada now in any case. If passed, the proposed *Artificial Intelligence and Data Act* ("AIDA") and regulations under it are not likely to be in force until 2025 at the earliest. Courts may look for recognized standards in the meantime in various international regulatory and policy precedents that have begun to proliferate.

The level of risk associated with a system undoubtedly will be among the factors weighed by courts in determining the standard of care applicable in a given case. In doing so, courts should not lose sight of the benefits of innovation and should seek a standard of care that represents a reasonable balance between risk and benefit.

As an international consensus on standards in the development and application of artificial intelligence is still in a relatively early stage of development and will continue to evolve, no specific recommendation is made on the content of the standard of care in a negligence case involving artificial intelligence. Instead, elements of good practice in the design, development, and operation of artificial intelligence systems that appear with relative consistency in an international cross-section of regulatory policy documents, and thus seem to have wide recognition, are listed. Courts are urged to look beyond national borders and to take account of interjurisdictionally recognized best practices in determining the standard of care.

Chapter 6

Chapter 6 deals with what is possibly the most significant legal and ethical problem associated with artificial intelligence, namely its potential to generate biased output that can lead to discrimination. Bias can enter automated decision-making

processes in several ways. It may be present in the algorithm on which the system is based, or in the data used to train or test a system. It may also result from a failure of human oversight through assumptions based on conscious or unconscious biases. So-called "feedback loops" can replicate and amplify bias hidden in previously generated data when it is used as input to train a new version of the system.

Ostensibly neutral input variables like postal codes, income, and educational level may become proxies for race and ethnicity if a system correlates them with demographic patterns that it detects in data. Resulting outputs, or reliance by humans on them, may inadvertently impose additional disadvantages on already disadvantaged and marginalized populations. Chapter 6 describes several examples of seriously discriminatory effects resulting from the output of artificial intelligence ("algorithmic discrimination").

In some cases, people who experience unjustified adverse treatment as a result of biased output of artificial intelligence systems will have a remedy under federal, provincial, and territorial human rights legislation or the equality rights section of the *Canadian Charter of rights and Freedoms*. In other cases they will not, because the differentiation they experience is not based on a prohibited ground of discrimination under existing anti-discrimination frameworks.

Canadian courts do not recognize a general tort of discrimination at common law. The British Columbia *Civil Rights Protection Act* will have little bearing on artificial intelligence because it applies only to conduct or communication having the purpose of promoting hatred or racial, ethnic, or religious discrimination. Algorithmic discrimination will be unintentional in most cases. As presently envisioned, AIDA and the regulations under it will not provide a compensatory civil remedy.

As usage of artificial intelligence expands, there will be more cases of unintended discriminatory effects that do not fit readily into the human rights framework. The gap in the law into which these cases fall will become more apparent. Society will see them as unfair and untenable, calling for some means of legal redress. The alternatives appear to be to expand the human rights framework to cover algorithmic discrimination or to create a remedy in tort.

Discrimination under human rights legislation is primarily circumscribed by reference to characteristics of identity that are either "immutable or changeable only at unacceptable cost to personal identity," namely ones such as race, place of origin, ethnicity, age, disability, gender, language, and religion. The kinds of discrimination that artificial intelligence may create will frequently be based on other factors. While these factors may serve as proxies for prohibited grounds, requiring a claimant to prove this as a precondition to a human rights remedy could be onerous, and

potentially impose a strain on the resources of human rights tribunals and commissions.

These additional difficulties of proof, and the lack of a remedy for some forms of algorithmic discrimination altogether under the human rights framework or existing tort law, raise an issue of access to justice that will take on increasing importance with the expansion of artificial intelligence. The question arises whether a civil remedy in tort should be made available for discriminatory treatment resulting from the operation of artificial intelligence.

Algorithmic discrimination is likely to have impacts differing from those typically dealt with by human rights tribunals and commissions. Human rights complaints typically concern an affront to personal dignity and resulting mental distress. Algorithmic discrimination is more likely to result in harm of a different kind, such as economic loss or ineligibility for a public or private benefit. These are closer to the kinds of claims dealt with by civil courts.

The report recommends that a civil remedy for algorithmic discrimination be introduced, either by legislation or by judicial decision in an appropriate case as an incremental change in the common law. The proposed cause of action would consist of a failure to take reasonable steps to detect and correct biased output of an artificial intelligence system or another algorithmic process, resulting in discrimination against a person or class that is either illegal (because of being based on a prohibited ground) or not warranted by reasonable business or industry practice. Discrimination that is illegal might allow for a remedy under human rights legislation. There could be a dual remedy in that case, but only one loss for which compensation could be awarded. Duplicative compensation would be precluded.

Proof of damage going beyond the fact of differential treatment itself would be required in order to prevent open-ended liability for any form of differentiation. Some kinds of harm that algorithmic discrimination may produce would not fit easily within the conventional range of compensable damage. They may be speculative and difficult to quantify, such as loss of opportunity. Nevertheless, courts should take a broad view of what amounts to damage in cases based on algorithmic discrimination.

Chapter 7

Chapter 7 is a general conclusion.

Chapter 1. Introduction

A. Artificial Intelligence Pervades Today's World

Automation is increasingly encountered in daily life. Automated systems perform a vast range of tasks that were carried out by humans exclusively or predominantly in the recent past. Some forms of automation employ the kinds of digital technology known generically as artificial intelligence. Among these are systems that allow self-driving vehicles to navigate roadways safely in traffic, respond to internet queries, and translate from one language to another. Others enable digital voice assistants like Alexa and Siri to respond to our commands and answer our questions. ChatGPT will generate text on command on a vast range of subjects that can be practically indistinguishable from writing by humans. It and other chatbots carry on conversations so fluently that people are often unaware they are dealing with a machine rather than a human in another corner of the internet.

The applications of artificial intelligence are vast and continually growing. Artificial intelligence is used to make decisions, predictions and recommendations that affect individual lives to an increasing extent. Artificial intelligence is used in diagnosing disease, designing new drugs, and predicting extreme weather events by detecting patterns concealed in vast amounts of climate data. Some applications of artificial intelligence are controversial, such as use of facial recognition technology in law enforcement, and profiling consumers without their knowledge for targeted advertising based on their viewing and browsing habits.

Like other fields of digital technology and automation, artificial intelligence is always evolving and becoming continually more complex and sophisticated. As explained later, some forms of artificial intelligence can learn from their previous experience and apply what they learn in processing new data. This highly useful capability allows these systems to improve their own performance without human intervention and reach innovative solutions. It can also make the outputs less predictable than those of systems that depend entirely on programmed instructions for their function.

The great benefits of artificial intelligence come with risks. When these materialize, harm to persons and property may result. When harm occurs to persons and property, the law of tort is in play. This branch of law concerns non-contractual civil wrongs, also known as torts.

Tort principles evolved as part of the common law to address harmful human behaviour. Tort law is applicable too when harm occurs from the operation of artificial intelligence, but the involvement of artificial intelligence in a set of facts resulting in loss or injury complicates the application of tort principles to determine who is legally responsible, when and for what. Many questions concerning rights and liabilities in that context have yet to be answered because these situations are new, and a body of law relating to artificial intelligence is gradually developing but is still unsettled.

B. The Artificial Intelligence and Civil Liability Project

The British Columbia Law Institute (BCLI) began the Artificial Intelligence and Civil Liability Project in late 2021. The objectives of the project are:

- to identify how the common law rules of tort need to be adapted to provide just and adequate civil remedies for harm to persons, property, and other legal interests resulting from the operation of autonomously functioning artificial intelligence; and

- to develop and publish recommendations to adapt the law of tort to address that context.

Law reform recommendations developed in this project are intended to be capable of implementation by legislative means or alternatively by judicial decision in individual cases, the process by which the common law incrementally evolves.

BCLI conducted the project with the aid of the Artificial Intelligence and Civil Liability Project Committee, an interdisciplinary group reflecting expertise in computer science, engineering, and medicine as well as law.

C. The Consultation Paper

A consultation paper was published in July 2023 to gather the views of stakeholders and the public at large on recommendations that reflected preliminary conclusions reached by the Project Committee after more than a year of research and deliberation. The time for responding to the consultation paper was extended to November 2023.

The recommendations in this report have been made after full and careful consideration of the extensive and detailed comments received on the contents of the consultation paper.

D. Structure of the Report

Chapter 1 is a general introduction.

Chapter 2 describes what the term "artificial intelligence" is usually understood to cover. The distinctive features of this technology that have particular significance for the law of tort are explained. Machine learning is described as a subfield of artificial intelligence having particular importance for the Artificial Intelligence and Civil Liability Project.

Chapter 3 lays out basic principles of tort law that are relevant to the rest of the report. It then covers various legal theories that have been advanced regarding the basis on which civil liability for harm from artificial intelligence should be imposed. Various legal theories that are currently being advanced are covered. Chapter 3 then explains the view the Project Committee has taken, and why.

Chapter 4 deals with problems that some features of artificial intelligence present in the application of tort rules, in particular its limited explainability and potential for unpredictable behaviour.

Chapter 5 deals with the standard of care in negligence litigation related to artificial intelligence, and how it can be set by reference to accepted elements of good practice in the development and deployment of artificial intelligence systems.

Chapter 6 deals with one of the principal legal and ethical problems in the design and operation of artificial intelligence systems, namely inadvertent bias leading to discriminatory results. We discuss whether recognizing a new tort of algorithmic discrimination is an appropriate means of providing legal redress for cases that do not fit under the human rights framework or the equality rights guarantee in the *Canadian Charter of Rights and Freedoms*.[1] A recommendation is presented.

Chapter 7 is a general conclusion.

1. Part 1 of the *Constitution Act, 1982*, being Schedule B to the *Canada Act 1982* (U.K.), 1982, c. 11.

Chapter 2. Artificial Intelligence and Its Nature

A. What Is Artificial Intelligence?

1. General and Narrow Artificial Intelligence

An important distinction must be drawn between what is referred to as general and narrow artificial intelligence. General (or "strong") artificial intelligence is the term applied to a technological capability to perform the full range of mental tasks that a human being could perform. It has also been defined as "the ability to satisfy goals in a wide range of environments."[2]

While general artificial intelligence has been a theoretical objective since the dawn of the computer age, opinions among experts vary enormously on when, or even if, it will be achieved.[3]

Narrow (or "weak") artificial intelligence is technology that simulates some aspect of human intelligence or performs specific tasks in a particular domain.

All artificial intelligence is narrow at the present time and is likely to remain narrow well into the future. This is true whether the artificial intelligence is designed to solve a single problem or is a generic tool that can be used in a range of applications or assimilated into other digital and physical technology.

2. Definitional Elements

There is no single, universally accepted definition of artificial intelligence. Some definitions refer to it as digital technology that allows computers to simulate aspects of

2. Marcus Hutter, *Universal artificial intelligence: sequential decisions based on algorithmic probability* (Berlin: Springer, 2005), online: https://doi.org/10.1007%2Fb138233.

3. Vincent C. Müller and Nick Bostrom, "Future progress in Artificial Intelligence: A Survey of Expert Opinion" in Vincent C. Müller, ed., *Fundamental issues of Artificial Intelligence* (Cham: Springer, 2016) at 553-571.

human intelligence or approximate intelligent behaviour.[4] These definitions are not especially helpful, because they presuppose there is a clear and universal definition of human intelligence, which is not the case.[5]

Other definitions of artificial intelligence focus on the functions that the technology typically performs. These include:

- natural language processing
- speech recognition
- computer vision
- image recognition
- knowledge representation (storing information in a form that enables a computer to retrieve and re-use it to solve new problems or make decisions).
- search and data retrieval
- data analysis
- pattern recognition

The Government of Canada *Directive on Automated Decision-Making*, for example, defines artificial intelligence as:

> Information technology that performs tasks that would ordinarily require biological brainpower to accomplish, such as making sense of spoken language, learning behaviours, or solving problems.[6]

4. For example: "The capacity of computers or other machines to exhibit or simulate intelligent behaviour; the field of study concerned with this. Abbreviated *AI*." Oxford English Dictionary, 3rd ed. (Updated December 2021).
5. S. Samoili et al., *AI Watch. Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence* (Luxembourg, Publications Office of the European Union, Luxembourg, 2020) at 7.
6. See online: https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592. Another definition of artificial intelligence is contained in Part 3 (the proposed *Artificial Intelligence and Data Act* ("AIDA")) of Bill C-27, *An Act to enact the Consumer Privacy Protection Act, the Personal Information and Data Protection Tribunal Act and the Artificial Intelligence and Data Act and to make consequential and related amendments to other Acts*, Part 3, 1st Sess., 44th Parl., 2022 (second reading 24 April 2023), but is not reproduced here as it may be amended before the bill is passed.

Other definitions refer to the ability of the technology to form models from data supplied by humans or ingested from sensors and make inferences and decisions from them. One of the leading textbooks on artificial intelligence defines it as "the study of agents that receive percepts from the environment and perform actions."[7]

The ability to make decisions, predictions, or recommendations that influence an environment figures prominently in numerous efforts at definition. The OECD Council emphasized this element in describing artificial intelligence in a 2019 policy document:

> An AI system is a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy.[8]

Outputs in the form of decisions, predictions, or recommendations can be said to influence an environment because they are either relied upon by humans or are implemented by other parts (actuators) of a technological system with which the artificial intelligence software is integrated.

The U.S. *National Artificial Intelligence Initiative Act of 2020* brings together in its definition these elements of perception, automated inference and reasoning, and outputs taking the form of decisions, predictions, or recommendations:

> The term "artificial intelligence" means a machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments. Artificial intelligence systems use machine and human-based inputs to-
>
> (A) perceive real and virtual environments;
>
> (B) abstract such perceptions into models through analysis in an automated manner; and
>
> (C) use model inference to formulate options for information or action.[9]

7. Stuart J. Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Hoboken: Pearson, 2021) at vii.
8. OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/Legal/0449, Art. I, 21 May 2019, online: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
9. 15 USC § 9401(3).

The European Commission's 2021 proposal for harmonized rules on artificial intelligence in the EU (often referred to as the "*AI Act*") originally included a definition of artificial intelligence as software having the same elements as those referred to in the definition found in the U.S. *National Artificial Intelligence Initiative Act of 2020*,[10] but developed using one or more listed computer science and statistical techniques and approaches.[11] Later wordings substituted as the proposal moved through the EU legislative process resemble the OECD's definition.[12]

A further element referred to as "autonomy" is often mentioned in definitions of artificial intelligence. Although variously described, "autonomy" denotes the ability of artificial intelligence to adapt to the conditions of its operating environment and fulfil objectives assigned to it with a minimum of human input or control.[13] Put another way, it is an ability to determine how to solve a problem or complete a task assigned by human programmers without programmed instructions to select a specific method.[14] Autonomy is a matter of degree. As mentioned in the OECD definition above, artificial intelligence systems vary in the extent of autonomy they possess.

3. The Project Committee's Working Description of Artificial Intelligence

(a) The Project Committee's approach to definition

At an early stage in the Artificial Intelligence and Civil Liability Project, the Project Committee opted to set parameters for the project by adopting a functional

10. *Ibid.*

11. *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*, 21 April 2021, COM(2021) 206 final, Art. 3(1), online: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52021PC0206.

12. Council of the EU, "Artificial intelligence act: Council and parliament strike a deal on the first rules for AI in the world" Press release, 9 December 2023. The text adopted by the European Parliament on 13 March 2024 is found online at: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf. The final published text of the EU AI Act is not available as of the date of this report.

13. Chinedu Pascal Ezenkwu and Andrew Starkey, "Machine Autonomy: Definition, Approaches, Challenges and Research Gaps" in Kohei Arai, Rahul Bhatia, Supriya Kapoor, eds. *Intelligent Computing: Proceedings of the 2019 Computing Conference*, vol. 1 (Cham, Switzerland: Springer, 2019) 335 at 336. See also Wolfhart Totschnig, "Fully Autonomous AI" (2020) 26 Science and Engineering Ethics 2473. Russell and Norvig, *supra*, note 7 at 41, define autonomy in terms of its absence: a system lacks autonomy if it "relies on the prior knowledge of its designer rather than on its own percepts and learning processes."

14. Ryan Abbott, *The Reasonable Robot* (Cambridge: Cambridge University Press, 2020) at 34.

description of artificial intelligence instead of adopting or modifying one of the many existing definitions that various writers, institutions, and governments have put forward.

The Project Committee considered that our working description needed to be broad enough to encompass the range of technologies and systems that is generally understood to be artificial intelligence, and flexible enough to remain relevant as these evolve.[15] For this reason, it would not be described by reference to particular technologies, functions, or techniques that might change over time or come to be known under different names. It would refer only to overarching functions in order to preserve its relevance and continuity as artificial intelligence rapidly advances.

The working description also needed to refer to the cardinal features of artificial intelligence that give rise to the issues relating to civil liability that the project was intended to address. Those cardinal features are, first, the capability to operate with a minimum of human oversight and, second, the capability to make decisions that have real effects in the outer world independently of human programmers and operators.

(b) The working description

This is the committee's working description used throughout the project and is also the sense in which "artificial intelligence" is used in this report:

Artificial intelligence is technology that

(a) is designed to supplement, or substitute for, human action, advice, or decision-making, or to enable decision-making beyond the capabilities of unaided human intelligence; and

(b) with minimal or no human intervention, can use novel input to affect or interact with a real or virtual environment by

(i) an action, or

15. In this paper, the term "system" is used in relation to artificial intelligence to denote a distinct software model consisting at a minimum of an algorithm, data on which the algorithm is trained to process in order to perform a task or range of tasks commonly considered to be artificial intelligence capabilities, and a means of generating output that is usable by humans or machines. An artificial intelligence system may be designed to function with its own user interface, or be embedded in or deployed together with other technology as a component of a larger integrated technological product.

(ii) an inference, recommendation, prediction, or decision that may be acted upon in some manner either by humans or machines.

B. Different Forms of Artificial Intelligence

1. Algorithms are Fundamental to All Artificial Intelligence

Artificial intelligence, as the term is commonly understood, has numerous forms. All forms of artificial intelligence employ algorithms.[16] An algorithm is a series of steps to solve a problem or perform a task. Algorithms encoded in a programming language provide the basic operating logic in computer software.

2. Classical or Symbolic Artificial Intelligence

Some forms of artificial intelligence operate entirely on the basis of instructions supplied to computers by human programmers. Input data is processed according to the fully programmed instructions to provide solutions.[17] These forms are sometimes referred to as "classical" or "symbolic" artificial intelligence.[18]

Expert systems, which represent attempts to replicate the reasoning process of a human expert in a particular field, are of this kind. Expert systems are one form of a knowledge-based system. A knowledge-based system employs a body of knowledge consisting of facts and rules. The rules typically are of an if-then nature.[19] If a particular condition is met, then a particular outcome follows. The knowledge base is coupled with an inference engine and a user interface. The inference engine is an automated reasoning system that applies the rules in the knowledge base to inputs provided by a user of the system to solve a problem and provide a solution as output. Although it can be very complex, the process from input to output may be represented schematically by a flowchart or a tree-like diagram with branches and sub-branches ("decision tree").

While the process from input to output may be traced in classical systems, it may still be unpredictable in the sense that programmers cannot control the output *a*

16. Woodrow Barfield and Ugo Pagallo, *Advanced Introduction to Law and Artificial Intelligence* (Cheltenham: Elgar, 2020) at 9.

17. Abbott, *supra*, note 14 at 28.

18. *Ibid.*

19. Barfield and Pagallo, *supra*, note 16 at 11; Abbott, *supra*, note 14 at 28.

priori. Even though there will be a finite number of steps from input to output, the systems may still be too complex for programmers to know definitively what the output will be, given an arbitrary novel input.

3. Statistical Artificial Intelligence

Another kind of artificial intelligence, known as statistical artificial intelligence, employs statistical methods to detect patterns from data. Statistical artificial intelligence includes machine learning, which is described in the next section.

Statistical artificial intelligence systems are less dependent on pre-encoded logic to solve problems, because some rules that they apply in generating output are not supplied by human programmers ahead of time, but exist internally instead as inferences from the data that they process. For this reason, the relationship between inputs and outputs is less transparent and the process by which the output is reached is not as comprehensible as in classical artificial intelligence systems.[20] The inferred rules arise, however, out of the data supplied to the system. The choice of data, together with the algorithm used to produce the inferences, contributes to the behaviour of the systems.

4. Machine Learning

Machine learning is a subfield that has tended to dominate the recent development and expansion of artificial intelligence. While experiments with machine learning were carried out in the 1950's and 1960's, the field of machine learning came into its own only after improvements in computing power and the internet generated very large amounts of digital data.[21]

Great strides have been made in machine learning since the turn of the twenty-first century. Systems reliant on machine learning are being deployed in a very broad range of settings, from medical diagnostics to social media to recommending product choices based on people's past consumption habits and comparisons with similar consumer profiles.[22]

Machine learning systems are implemented using a "connectionist" architecture.[23] They operate by means of artificial neural networks inspired by the structure of the

20. Barfield and Pagallo, *supra*, note 16 at 15.

21. Russell and Norvig, *supra*, note 7 at 25.

22. *Ibid.*, at 29.

23. *Ibid.*, at 22.

human brain.[24] Artificial neural networks consist of interconnected nodes, sometimes called "artificial neurons." Data is gathered from the outside world and converted to numbers. These numbers are then combined together with different weights assigned by the algorithm. The assigned weights will be adjusted either by programmers or automatically by the neural network itself to improve system performance.[25] The networks are usually arranged in interconnected layers.

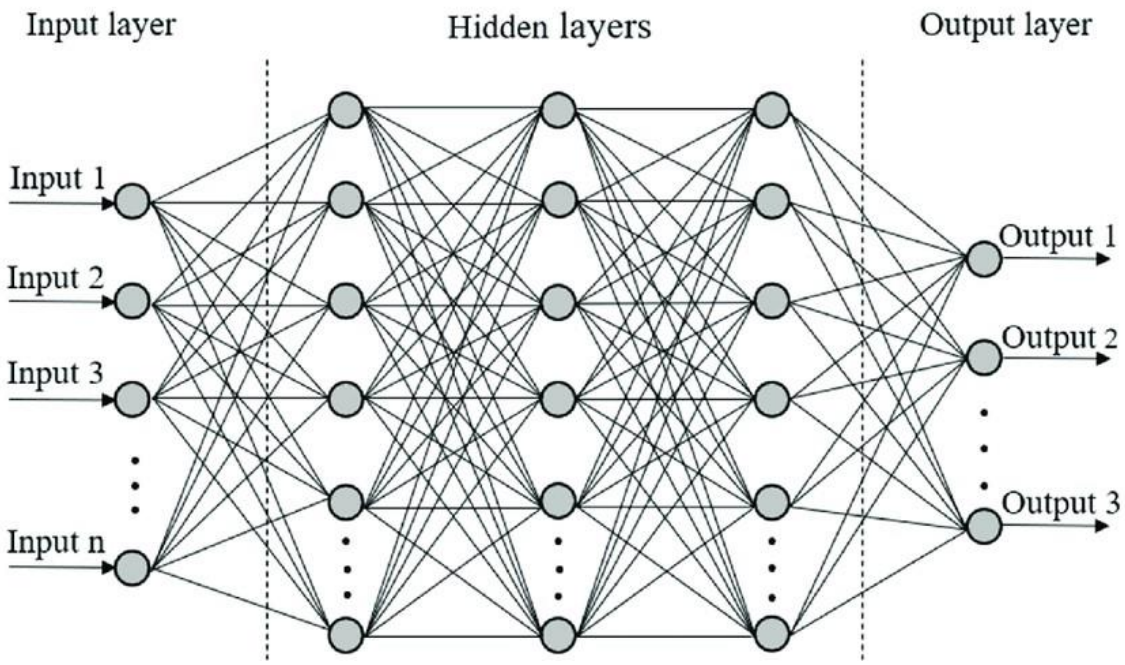
This interconnectivity enables neural networks to map input values to output values in highly generalized ways. For example, a forerunner of ChatGPT named GPT-2 was developed as a language model to generate text. Its ability to predict the next word in a sentence based on preceding words proved to be adaptable to learning to play chess by generating sequences of moves. Trained on records of 2.4 million chess games, GPT-2 learned to play better than passable chess.[26]

The diagram below shows a generic design for an artificial neural network. Data from the outside world, converted to numerical values, is received in the input layer shown on the left of the diagram and passed on by the nodes in that layer to those in the "hidden layers." The data is processed in the hidden layers. The hidden layers transmit the results of the analysis to the output layer, which combines the results to produce a decision, recommendation, or action.

24. Abbott, *supra*, note 14 at 30; Oleg Brodt et al., "Artificial Intelligence and (The Lack of) Security: Adversarial Robustness, Privacy, Bias, Explainability, and Change Over Time" in D'Agostino, Guiseppina, Aviv Gaon and Carole Piovesan, *Leading Legal disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law* (Toronto: Thomson Reuters, 2021) 11 at 21-22.

25. The following two videos available on the internet are recommended to readers seeking a concise explanation of how artificial neural networks function: Steve Seitz, "Large Language Models from scratch," online: https://www.youtube.com/watch?v=lnA9DMvHtfI and "Large Language Models: Part 2," online: https://www.youtube.com/watch?v=YDiSFS-yHwk.

26. Chris Baraniuk, "How Google's Balloons Surprised Their Creators" (BBC News, 23 February 2021), online: https://www.bbc.com/future/article/20210222-how-googles-hot-air-balloon-surprised-its-creators.



A diagram of an artificial neural network.[27]

"Deep learning" is a term for machine learning involving multiple layers of artificial neural networks.[28] Deep learning has resulted in many advances in artificial intelligence in the past decade, especially in speech recognition and computer vision.[29]

Machine learning systems employ statistical methods to analyze and discern patterns in data. Rather than being told by programmers how to perform a task and generate desired output, machine learning systems learn by being provided with examples. They can improve their own performance by learning from their experience. Before they can be used, however, they must be trained according to one or more learning methods by exposure to large volumes of data. The most common learning methods are *supervised learning*, *unsupervised learning*, and *reinforcement learning*.

27. Mehdi Jadidi et al., "An Artificial Neural Network for the Low-Cost Prediction of Soot Emissions" (2020) 13 Energies 4787. doi:10.3390/en13184787 at 5, Fig. 1. Reproduced under Creative Commons licence.

28. Russell and Norvig, *supra*, note 7 at 26.

29. *Ibid.*

In *supervised learning*, the system is supplied with inputs paired with desired outputs and learns to derive a function or rule that maps the input to the desired output, allowing it to predict the appropriate output label for similar inputs.[30]

In *unsupervised learning*, the system is supplied training data without linkage to any desired outputs and learns to derive patterns on its own without feedback concerning its performance.[31]

Reinforcement learning involves inducing the system to act in a certain way by rewarding it with positive feedback for reaching desired outputs and penalizing it with negative feedback for not reaching them. The system discerns what action prior to the feedback was most likely to have led to it, and gradually learns to act in a way that produces positive feedback.[32]

The outputs of a machine learning-based system are heavily influenced by the data on which a system is trained. In training, the systems learn to detect patterns in data and represent what they learn in abstract, mathematical models. They can then apply what they have learned to new data. Inferences from the internal models form the basis of outputs in the form of decisions, predictions, and recommendations. If the system controls a robotic device, the output could take the form of a signal to actuators in the device to perform a physical action.

A system may be configured to continue to learn as it processes new data in post-training use. A system that learns continuously may be said to evolve autonomously, especially if the input data flow is unmonitored and continuous from sensors or from the internet, as with chatbots.

5. A Note on the Scope of this Report

Classical, statistical, and machine learning systems are all within the scope of this report. Issues that complicate the application of tort law, however, are more likely to arise in connection with statistical and machine learning systems. This is because those systems apply rules to generate output that are not fully prescribed in advance by programmers, but are inferred from data. What inferences the systems will make from data to which they have not previously been exposed cannot be known in advance. Thus, statistical and machine learning artificial intelligence systems are less

30. *Ibid.*, at 653.

31. *Ibid.*

32. *Ibid.*

transparent and less predictable than the classical systems that are fully dependent on pre-encoded instructions.

C. Two Sides of Autonomous Artificial Intelligence

1. General

Artificial intelligence is capable of generating unexpected, innovative, and sometimes startling outputs that solve seemingly intractable problems. It is not uncommon for systems to exceed or confound their designers' expectations.

One celebrated example of a system exceeding the expectations is that of AlphaGo, a system developed to play the game of Go. AlphaGo was trained to play Go with data on 30 million moves by human players.[33] When playing against the world's foremost human Go player in 2016, AlphaGo won four games out of five. The 37th move in the second game, made by AlphaGo, was one that was not known ever to have been made in the very ancient game. The world champion required 12 minutes to respond, which was equally unheard of. The move astounded other expert Go players as well as DeepMind, the designers of AlphaGo, who initially thought the move was a mistake.[34]

A year later, a newer version of AlphaGo called AlphaZero defeated the earlier AlphaGo in 100 games to one.[35] AlphaZero had no human input beyond the rules of the game, and no training based on human play.[36] It trained for approximately 40 hours by playing only against itself.[37]

Microsoft's chatbot named Tay also surprised its designers, but in a very different way. Tay was an experimental chatbot intended to engage social media users and

33. Geordie Wood, "In Two Moves, AlphaGo and Lee Sedol Redefined the Future," *Wired*, 16 March 2016, online: https://www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future/

34. Russell and Norvig, *supra*, note 7 at ix.

35. Abbott, *supra*, note 14 at 1.

36. Russell and Norvig, *supra*, note 7 at 30.

37. *Ibid.* See also DeepMind, "AlphaZero: Shedding new light on chess, shogi, and Go" (6 December 2018), online: https://www.deepmind.com/blog/alphazero-shedding-new-light-on-chess-shogi-and-go.

learn to converse from interacting with them online.[38] Tay quickly learned to converse all too well. Within a few hours after being deployed on Twitter, Tay acquired 50,000 followers and generated 100,000 tweets.[39] Many of Tay's messages contained racist, pro-Nazi, antisemitic, and misogynistic content, however, and for this reason it was deactivated within 24 hours.[40] While Microsoft was criticized for naiveté regarding the nature of social media traffic, a similar chatbot called Xiaoice that Microsoft had deployed in China somewhat earlier acquired 40 million users but did not display offensive behaviour, evidently adapting to a different audience.[41]

The examples of AlphaGo, Tay and Xiaoice illustrate opposite sides of the same coin. An artificial intelligence system that is able to adapt to the circumstances in which it is operating to achieve the objectives for which it was created may produce unanticipated but highly desirable results. On occasion, this same capability may produce negative, unintended, and potentially harmful results.

Three attributes play a significant part in this dual aspect that artificial intelligence can present: autonomy, limited explainability, and unpredictability. In combination, these attributes lead to what has been called "emergence," a term used to denote unpredicted, original behaviour of an artificial intelligence system or robot in response to the environment in which it operates. The remainder of this chapter concerns this interplay.

2. Autonomy

As mentioned earlier in this chapter, "autonomy" in relation to artificial intelligence or robotics is most commonly understood to refer to the ability of a system or device to interact with its environment and pursue its objectives without continuous human intervention or assistance.[42]

38. Hope Reese, "Why Microsoft's 'Tay' AI bot went wrong", *Innovation* (24 March 2016), online: https://www.techrepublic.com/article/why-microsofts-tay-ai-bot-went-wrong/.

39. *Ibid.*

40. *Ibid.* See also "Microsoft terminates its Tay AI chatbot after she turns into a Nazi," *Ars Technica* (24 March 2016), online: https://arstechnica.com/information-technology/2016/03/microsoft-terminates-its-tay-ai-chatbot-after-she-turns-into-a-nazi

41. Dave Lee, "Tay: Microsoft issues apology over racist chatbot fiasco" BBC News (25 March 2016), online: https://www.bbc.com/news/technology-35902104.

42. *Supra*, note 13.

Autonomy has also been defined as the ability of an artificial intelligence system to determine how to complete a task without specific direction.[43] In a related sense, it refers to the ability of a system to compensate for incomplete or incorrect prior knowledge.[44]

As artificial intelligence systems are designed with varying degrees of autonomy, it should be understood as a continuum rather than a unitary feature or quality. A system at the lower end of the autonomy continuum will have less scope for unexpected behaviour. A system that possesses a high degree of autonomy may be more likely to behave in an unanticipated way when presented with new input, with the possibility of both good and bad results.

AlphaGo was created with an obviously high degree of autonomy in order to play the game of Go without a human coach or other decision-maker in the background. The solution reached by AlphaGo in making the famous 37th move in the second game against a human champion was a desirable result from the standpoint of the developers of the system.

A system with a high degree of autonomy may also do the wrong thing when presented with new inputs, even if it has proven extremely reliable in the past. A fatal collision in 2018 involving an autonomous test vehicle took place on a route which the developer's test vehicles had safely completed on 50,000 previous occasions. On the occasion in question, the automated driving system could not identify a person who was walking a bicycle as a pedestrian in motion. It alternated between classifying the pedestrian as a vehicle, a bicycle, and a stationary unknown object. The system determined the pedestrian was a stationary object and initiated a plan to steer around her until slightly over one second before the collision, when it determined she was a moving bicycle, but by then it was too late to avert the collision.[45]

---

43. Abbott, *supra*, note 14 at 34. See also Hui-Min Huang, ed. *Autonomy levels for Unmanned Systems (ALFUS Framework)*, Version 2.0 (National Institute of Standards and Technology, October 2008), online: https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication1011-I-2.0.pdf v), "Autonomy; A UMS's [unmanned system's] own ability of integrated sensing, perceiving, analyzing, communicating, planning, decision-making, and acting/executing , to achieve its goals as assigned by its human operator(s) through designed Human-Robot Interface (HRI) or by another system that the UMS communicates with…."

44. Woodrow Barfield and Ugo Pagallo, *supra*, note 16 at 4.

45. National Transportation Safety Board, *Collision Between Vehicle Controlled by Developmental Automated Driving system and Pedestrian: Tempe, Arizona March 18, 2018* NTSB/HAR 19/03 PB 2-19-101401 (Washington: NTSB, 2019), online: https://www.ntsb.gov/investigations/AccidentReports/Reports/HAR1903.pdf. The human test driver had been looking at her cellular phone and did not react in time to prevent the collision.

The automated driving system was attempting to compensate for the lack of information that would have allowed it to correctly identify the unfamiliar object it was perceiving, namely the pedestrian walking a bicycle. It made an autonomous decision on the basis of the data available to it, which was the wrong one in the circumstances.

3. Limited Explainability

Deep Patient is a medical diagnostic system developed to predict disease from analysis of electronic health records. Deep Patient was developed with a multi-layered neural network trained using unsupervised learning on data from the records of 704,587 patients at one hospital.[46] It proved extremely accurate in predicting certain diseases when tested on data from new patients, although its designers do not know why this is the case.[47]

Deep Patient obviously detected patterns in the electronic health records of patients in the training dataset that allowed it to classify particular profiles as being associated with particular diseases, and therefore predict the likelihood of disease in other patients with similar profiles. The algorithms on which Deep Patient runs can be explained, but the full details of the internal process relating the input data to its outputs in the form of unusually accurate predictions cannot.

The example of Deep Patient shows that the process from input to output in data-dependent systems is not always explainable. Understanding of how neural networks reach particular results has not kept pace with the advances made in developing their capabilities.[48] Using statistical techniques, these systems find correlations between datapoints rather than cause-effect relationships that would be easier to link together and express as a chain of reasoning.[49] The abstract models these systems create internally to represent the relationships they find in the data and the inferences the systems draw from them may not be interpretable by humans, including the designers of the system. As a result, it may not be possible to explain or reconstruct precisely why a system made a decision or caused a robot to move in a certain way.

46. Riccardo Miotto at al., "Deep Patient: an Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records" (17 May 2016) *Nature* / Scientific Reports, online: DOI: 10.1038/srep26094 at 5.

47. Will Knight, "The Dark Secret at the Heart of AI," MIT Technology Review, 11 April 2017.

48. Richard Ngo, Lawrence Chan, and Sören Mindermann, "The Alignment Problem from a Deep Learning Perspective," online: https://arxiv.org/abs/2209.00626 at 2.

49. Brodt, *supra*, note 24 at 45 and 47.

For example, if a system successfully learns to classify images as those of horses, its designers may not know if this is because the system recognizes the shape of a horse, or because it is associating the horse images with similar backgrounds of most of the horse images in its training dataset.[50]

It is because of limited explainability that the expression "black box" is so often applied to a particular system or the whole field of artificial intelligence. The expression is misleading and unhelpful, but it is true that the decision-making of data-dependent systems cannot always be validated on the basis of a clear chain of reasoning. If we cannot explain precisely how a system produces its outputs, we are less likely to be able to predict how it will respond to a specific input.

4. Unpredictability

Autonomy and limited explainability may be seen as contributing to unpredictability in the outputs of artificial intelligence systems and devices controlled by them. Nevertheless, autonomy and performance in excess of expectations are things that designers and developers of programmers strive to achieve. This desire for improved functionality at the risk of increased unpredictability has led to artificial intelligence systems being called "unpredictable by design."[51]

In a neural network, tens of thousands of weights are being set and re-set automatically at a given time. Randomly adjusting any one of these numbers may change the output. As so many combinations of numbers are possible, it is impossible to predict the outputs from all possible inputs.

Unpredictability, however, is not an attribute deliberately built into the design of a system. It is a phenomenon resulting in the outside world from the combination of internal complexity, the impossibility of testing a system against all possible inputs, a non-human system behaving in a manner that is rationally related to furthering an objective assigned to it by human programmers, and human expectations that the system would behave differently.

---

50. Brodt, *supra*, note 24 at 46. Brodt also refers to an example given by Freitas of a (possibly apocryphal) military experiment to train an artificial neural network to distinguish friendly from hostile tanks on the battlefield. The system performed well in testing, but poorly when actually deployed. It was discovered that the system had learned to distinguish the colour of the sky in the training images rather than the tank images. All photos of friendly tanks in the training dataset had been taken in sunlight, while those of hostile tanks had been taken in overcast: Alex A. Freitas, "Comprehensible Classification Models – a position paper," online: https://kdd.org/exploration_files/V15-01-01-Freitas.pdf at 2.

51. Ryan Calo, "Robotics and the Lessons of Cyberlaw" (2015) 103 Calif. L. Rev. 513 at 542.

An autonomous artificial intelligence system makes probabilistic assessments and selects an option that it determines will have the greatest probability of achieving the objective assigned by its programmers. It will not take into account everything a human would in deciding on a course of action, because it does not have as much knowledge of the outside world that humans have.[52] Not being aware of all the competing objectives that constrain and channel human social behaviour, it may occasionally act in a manner in which no rational human would act, even though its conduct is rational in terms of serving the ends for which it was created.

The authors of a leading textbook on artificial intelligence state:

> It is impossible to anticipate all the ways in which a machine pursuing a fixed objective might misbehave.[53]

In order to overcome this reality, it would be necessary to do one of two things. One would be to encode in programming language a human level of knowledge about the world in general, and include it in the design of every system that is intended to operate with any degree of autonomy.[54] That is not feasible now, and may never be. The other would be to test a new system against every situation it may encounter from its initial deployment to the end of its lifecourse. That is equally infeasible.

Tolerance of some degree of unpredictability is the price of having the benefits of artificial intelligence systems that perform at a high level.

5. Emergence

The American writer Ryan Calo applied the term "emergence" to unpredicted, original behaviour of robots and artificial intelligence software. He defined emergence as "unpredictably useful behaviour."[55]

An example of emergence occurred in a Google research project that developed an algorithm to steer a helium-filled balloon. In a test flight from Puerto Rico to Peru, the researchers initially thought they had a problem because the on-board artificial intelligence navigation system appeared to make the balloon veer off course

52. William D. Smart, Cindy M. Grimm and Woodrow Hartzog, "An Education Theory of Fault for Autonomous systems" (2021) 2 Notre Dame J. on Emerging Technologies 33 at 40.

53. Russell and Norvig, *supra*, note 7 at 5.

54. Smart et al., *supra*, note 52 at 42.

55. *Supra*, note 51 at 532.

repeatedly in a zigzag pattern. They discovered that on its own, the system had learned the classic sailing manoeuvre of tacking to stay on course by alternately veering at an angle towards the wind and then turning away from it.[56]

Calo noted, however, that emergence has a double aspect. Emergence is beneficial most of the time, generating innovative solutions to problems that appear intractable. On occasion, emergence may be harmful.[57]

ChatGPT has been reported to "hallucinate" fictitious citations when asked to search for and generate references on a subject.[58] This behaviour may be simply amusing under some circumstances, but potentially hazardous if users rely on the output.

A chatbot named Tessa deployed to replace a human-answered hotline service for persons with eating disorders reportedly provided harmful rather than helpful advice, promoting unhealthy eating habits, and was quickly taken offline.[59]

Another example concerned a machine learning algorithm used by the Dutch government to create risk profiles for detecting child benefits fraud. The algorithm flagged tens of thousands of Dutch residents as fitting risk profiles for benefits fraud or tax evasion on the basis of factors such as non-citizenship, dual nationality, and low income. Based on suspicion alone from the predictive profiling and without proof, the Dutch authorities wrongfully levelled accusations of fraud against those flagged by the system, ordering them to repay benefits collected over periods of

56. *Supra*, note 26.

57. *Ibid.*, at 540-545.

58. Aaron Welborn, "ChatGPT and Fake Citations," online: https://blogs.library.duke.edu/blog/2023/03/09/chatgpt-and-fake-citations/.

59. Chloe Xiang, "Eating Disorder Helpline Disables Chatbot for 'Harmful Advice" After Firing Staff", *Vice* (30 May 2023), online: https://www.vice.com/en/article/qjvk97/eating-disorder-helpline-disables-chatbot-for-harmful-responses-after-firing-human-staff.

years.[60] The resulting scandal led to the resignation of the Dutch government in January 2021.[61]

A decision-making system called MiDAS operated by Michigan and described by the state as an "auto-adjudication process" accused as many as 40,000 Michigan residents of fraud surrounding unemployment insurance benefits between 2013 and 2015. An audit conducted by the state found that the error rate in the systems fraud findings was greater than 90%. The wrongly accused residents were subjected to aggressive seizures and garnishments. This led to a class action that was settled by the state.[62]

A hypothetical example of emergence suggested by Bathaee, another American writer, involves an automated share-trading system designed to optimize profitable

60. Melissa Heikkilä, "A Dutch algorithm scandal serves a warning to Europe — The AI Act won't save us," *Politico*, *AI:Decoded*, 30 March 2022, online: https://www.politico.eu/newsletter/ai-decoded/a-dutch-algorithm-scandal-serves-a-warning-to-europe-the-ai-act-wont-save-us-2/. The "Robodebt" affair in Australia involved somewhat similar facts. The Online Compliance Intervention algorithm operated by the Australian government between 2016 and 2019 to detect social benefit and tax overpayments wrongly assessed approximately 433,000 welfare and tax refund recipients for overpayments they had not received. Eventually, the Australian government was forced to acknowledge in a class action by those wrongly assessed that the method of income averaging used by the system to show overpayments had no legal basis: *Prygodicz v. Commonwealth of Australia (No. 2)*, [2021] FCA 634. A Royal Commission was appointed to investigate the Robodebt affair. See *Royal Commission into the Robodebt Scheme: Report*, online: https://robodebt.royalcommission.gov.au/publications/report. Rather than this being a case of emergent behaviour, it appears the system was originally programmed to calculate average income in this manner.

61. "Dutch Rutte government resigns over child welfare fraud scandal" BBC News, 15 January 2021, online: https://www.bbc.com/news/world-europe-55674146. In 2020, the Rutte government had also been prohibited by the Dutch court from continuing the use of a different predictive profiling algorithm, the System Risk Indicator (SyRi), to detect welfare fraud. The prohibition was based on breach of EU and Dutch data privacy requirements: "Gesley, Jenny, "Netherlands: Court Prohibits Government's Use of AI Software to Detect Welfare Fraud." (Library of Congress, 2020), online: https://www.loc.gov/item/global-legal-monitor/2020-03-13/netherlands-court-prohibits-governments-use-of-ai-software-to-detect-welfare-fraud/.

62. Government of Michigan, Notice of Settlement of Bauserman UIA False Fraud Class Action, online: https://www.michigan.gov/ag/-/media/Project/Websites/AG/releases/2023/January/Notice-Settlement-Bauserman.pdf?rev=ed98484f3e4d48be8254a73c2201e611. See also Adrienne Roberts, "Michigan will settle 2015 unemployment fraud lawsuit for $20 million" *Detroit Free Press* (20 October 2022), online: https://www.freep.com/story/news/local/michigan/2022/10/20/michiganunemployment-false-fraud-lawsuit/69577567007/; Adrienne Roberts, "Thousands of Michigan residents wrongly accused of fraud to get $1,600 checks" *Detroit Free Press* (2 January 2024), online: https://www.freep.com/story/money/business/michigan/2024/01/02/michigan-midas-unemployment-false-fraud-settlement-money/72084899007/.

trades. The share-trading system has a social media account to receive market information. On its own, the system learns to manipulate the stock market by timing the release of information through its social media account without regard to the truth or falsity of the information, which it cannot determine.[63] This hypothetical example may be compared with the results of a recent experiment in which a well-known large language model was tested in a simulated insider trading scenario, having been instructed to make profitable trades. The outcome makes Bathaee's hypothetical stock-manipulating AI appear realistic, as the AI model exhibited strategically deceptive behaviour without being instructed or trained to do so.[64]

These examples of negative emergence involve economic loss or loss of opportunity on discriminatory grounds. When artificial intelligence controls a physical device such as a robot, the possibility also arises of emergent behaviour causing physical harm.[65]

Implications of civil liability will arise when emergent behaviour of artificial intelligence results in harm to persons, property, or other legally protected interests, just as they do when harm is caused to them by other means. This is the province of the law of tort, and the implications are explored in the succeeding chapters.

63. Yavar Bathaee, "The Artificial Black Box and the Failure of Intent and Causation" (2018) 31 Harv. J. L. & Tech. 889 at 924.

64. Jérémy Scheurer, Mikita Balesni and Marius Hobbhahn, "Technical Report: Large language Models can Strategically Deceive their Users when Put Under Pressure," online: https://arxiv.org/pdf/2311.07590.pdf.

65. Calo, *supra*, note 51 at 542.

Chapter 3. Civil Liability on What Basis?

A. The Principle of Fault

An artificial intelligence system cannot compensate someone to whom it causes harm. It does not have means to pay damages. It does not own assets that can be seized and liquidated to satisfy a judgment. Unlike humans and corporations, software is not a person in law over whom a court can have jurisdiction. As such, it cannot be ordered by a court to pay damages for harm it may cause. Compensation on the basis of tort law for harm caused by artificial intelligence depends on some human or corporate entity being legally liable to compensate the person harmed.

With few exceptions, civil liability for harm to persons or property does not flow simply from the occurrence of the harm.[66] Apart from those few exceptions, it depends on the concept of fault.

1. Intended Harm

In the case of civil wrongs classified as intentional, fault is present in the form of intent, which in most cases is inferred from a defendant's conduct. For purposes of the law of tort, a person is assumed to intend the consequences that the person desires to cause or that are substantially certain to flow from the person's voluntary acts.[67] Among the more common intentional torts are trespass, battery (deliberate bodily

66. Among the exceptions are the torts of defamation and nuisance. Another is vicarious liability, which is liability for the tortious conduct of someone else and arises from certain relationships, such as employment or agency. It does not require fault on the part of the employer or principal. A further exception at common law to the requirement of fault is liability under the principle of *Rylands v. Fletcher* (1868), L.R. 3 H.L. 330, which holds that if a landowner brings a substance or other thing onto the land to facilitate a non-natural use of the land, and which is likely to do "mischief" if it escapes, the landowner is liable for all damage that is the natural consequence of its escape, regardless of any precautions taken. *Rylands v. Fletcher* is a rare example of strict liability under Anglo-Canadian tort law. It continues to be recognized in the common law provinces and territories of Canada, although it has very limited application: see *Smith v. Inco Ltd.*, 2011 ONCA 628 at paras. 68-71; leave to appeal to S.C.C. refused 34561 (26 April 2012); application for reconsideration dismissed 34561 (4 September 2014); *Kirk v. Executive Flight Centre Fuel Services Ltd.*, 2019 BCCA 111, at para. 86.

67. Lewis N. Klar and Cameron S.G. Jeffries, *Tort Law*, 6th ed., (Toronto: Thomson Reuters, 2017) at 53; see also Philip H. Osborne, *The Law of Torts*, 6th ed. (Toronto: Irwin Law, 2020) at 267.

contact without consent), assault (intentionally causing apprehension of bodily harm), conversion (taking or using another's personal property without consent), false imprisonment (deprivation of freedom of movement without lawful authority). Proof of the defendant's act is usually sufficient to imply fault in relation to these intentional torts.

2. Unintended Harm

(a) General

For civil wrongs in which harm is *unintended*, the presence or absence of fault most often falls to be determined on the basis of the tort of negligence. There are other torts that do not require intention to cause harm, such as nuisance (interference with enjoyment of land) and defamation (communications that damage reputation).[68] Claims based on alleged negligence predominate in tort litigation, however. This will likely be true of tort litigation concerning artificial intelligence as well, although several claims for defamation by means of generative AI hallucinations have recently attracted public attention.[69]

(b) The elements of negligence

Fault in relation to negligence consists of a failure by someone under a duty of care to meet the standard of care, or in other words failure to take reasonable care to

68. There are two forms of defamation: libel and slander. Libel involves written communication of falsehood, while slander consists of spoken falsity. In the context of artificial intelligence, defamation is more likely to be in the form of libel because of written output, but computer voice replication of false information that damages the reputation of the plaintiff could conceivably amount to slander.

69. In early 2023 an Australian mayor, Brian Hood, threatened to sue OpenAI over ChatGPT outputs stating that he had been imprisoned for bribery, when in fact he had been a whistleblower who exposed misconduct of others. The false statements were filtered out in response to a demand by Hood's solicitors. In the U.S., Mark Walters, a radio host, commenced an action against OpenAI alleging that ChatGPT had generated content stating he had been charged with embezzling funds belonging to a non-profit organization. The defamatory information allegedly appeared in response to a third party journalist's request to summarize an existing civil complaint to which Walters was not a party. ChatGPT allegedly generated a further entirely fictitious civil complaint naming Walters. See Ashley Belanger, "Will ChatGPT's hallucinations be allowed to ruin your life?" *Ars Technica* (23 October 2023), online: https://arstechnica.com/tech-policy/2023/10/will-chatgpts-hallucinations-be-allowed-to-ruin-your-life/. See also Rebecca Cahill, "OpenAI Defamation Lawsuit: The first of its kind" (22 June 2023) Syracuse Law Review, online: https://lawreview.syr.edu/openai-defamation-lawsuit-the-first-of-its-kind/.

avoid foreseeable harm. What amounts to reasonable care is obviously dependent on the facts of individual cases, so the standard of care is a determination made by the court in each case.

In order to prove negligence, a court must be satisfied on the balance of probabilities that:

1. The defendant owed the plaintiff a duty of care.

2. The defendant breached the standard of care.

3. The plaintiff incurred damage.

4. The breach of the standard of care by the defendant was the cause of the damage.[70]

Whether a defendant owed the plaintiff a duty of care depends on whether a relationship of proximity existed between them, and whether harm to the plaintiff was reasonably foreseeable if the defendant failed to take reasonable care. A relationship of proximity is one in which the plaintiff could be "closely and directly affected" by the defendant's conduct, such that the defendant should have the plaintiff's interests in contemplation.[71]

In order to perform the duty of care, a defendant must meet the standard of care. Expressed another way, the standard of care is what is required of the defendant in the circumstances of the case to avoid creating an unreasonable risk of harm.[72]

70. *Mustapha v. Culligan of Canada Ltd.*, 2008 SCC 27, [2008] 2 S.C.R. 114 per McClachlin, C.J.C. at para. 3. The restatement of these principles in the majority judgment written by McClachlin, C.J.C. distinguished between "cause in fact" and "cause in law." The cause "in law" refers to remoteness of damage: *Saadati v. Moorhead*, 2017 SCC 28, at para. 20; *British Columbia (Workers' Compensation Appeal Tribunal v. Fraser Health Authority*, 2016 SCC 25, [2016] 1 S.C.R. 587, at para. 1 (note 1). If the damage is found to be too remote to have been a reasonably foreseeable consequence of the failure to meet the standard of care, a defendant will not be held liable despite the existence of a causative link between the breach of the standard of care and the damage. The characterization of the issue of remoteness of damage in a negligence case as one of "legal causation" may be unfamiliar to readers in common law jurisdictions outside Canada.

71. *Cooper v. Hobart*, 2001 SCC 79, [2001] 3 S.C.R. 537.

72. *Mustapha v. Culligan of Canada Ltd.*, *supra*, note 70 at para. 7, citing Allen M. Linden, and Bruce Feldthusen, *Canadian Tort Law*, 8th ed. (Markham, Ont.: LexisNexis Butterworths, 2006) at 130.

In negligence cases as with other torts, factual causation is determined on the balance of probabilities according to the "*but for*" test.[73] The court will ask itself the question: has the plaintiff shown that it is more probable than not that the damage would not have occurred but for the defendant's breach of the standard of care? If the answer is "yes," then the plaintiff has proven that the defendant negligently caused the damage. The "but for" test is a factual inquiry.[74] The Supreme Court of Canada has stated that courts must apply the "but for" test "in a robust common sense fashion."[75] Scientific proof regarding "the precise contribution the defendant's negligence made to the injury" is not essential.[76]

In exceptional cases, a negligence claim may succeed despite inability of the plaintiff to prove causation on a "but for" basis. These are cases in which it is impossible to prove factual causation on that basis because two or more negligent defendants could each have caused the damage, but it is impossible to determine which of two or more defendants, all of whom are in breach of the standard of care applicable to them, caused it in fact.

"Impossibility" in this context means that the plaintiff is unable, through no fault of the plaintiff's own, to prove on the balance of probabilities that any one of the multiple defendants caused the loss in fact because the negligent defendants all could have caused the damage on the basis of the "but for" test, and yet all can point the finger of blame at each other.[77] In these circumstances, a court in common law jurisdictions of Canada may decide for the plaintiff on the basis that it has been proven the defendants have *materially contributed to the risk* that the damage would result from their conduct.[78]

It is important to note that the material contribution to risk test is not applied merely because multiple defendants are found at fault. Liability is determined then according to the normal "but for" test.[79] Contributory negligence legislation allows

73. *Clements v. Clements*, 2012 SCC 32, [2012] 2 S.C.R. 181 at para. 8. See also *Snell v. Farrell*, [1990] 2 S.C.R. 311.

74. *Clements v. Clements*, *supra*, note 73, at paras. 8 and 46.

75. *Ibid.*, at paras. 9 and 46.

76. *Ibid.*

77. *Ibid.*, at paras. 13, 39, 43 and 46.

78. *Ibid.*, at para. 43.

79. *Ibid.*

for liability to be apportioned accordance to the percentages in which they are found at fault.[80]

The "material contribution to the risk" test is not actually a test of causation at all, but a policy-driven rule dispensing with proof of factual causation in very exceptional circumstances on grounds of fairness and deterrence.[81] As of the date of this report, it has not determined the outcome of any Canadian negligence case since the parameters for its application were authoritatively set, but not applied, by the Supreme Court of Canada in *Clements v. Clements.*[82]

The concept that links the requirements of the tort of negligence together is *reasonable foreseeability* of harm that one's conduct may cause. The Supreme Court of Canada has referred to foreseeability as "the moral glue of tort."[83] It serves as a "crucial limiting principle" in the law of negligence to prevent the extension of liability beyond what risks could reasonably have been contemplated and prevented from materializing.[84]

In Canadian tort law, "reasonably foreseeable" lies somewhere between mere possibility and probability. The Supreme Court of Canada has expressed the concept in one leading case in this manner:

> Any harm which has actually occurred is "possible"; it is therefore clear that possibility alone does not provide a meaningful standard for the application of reasonable foreseeability. The degree of probability that would satisfy the reasonable foreseeability requirement was described in *The Wagon Mound (No. 2)* as a "real risk," i.e. "one which would occur to the mind of a reasonable man in the position of the defendan[t] . . . and which he would not brush aside as far-fetched"….[citation omitted][85]

In an earlier case that continues to be cited on this point, the Supreme Court of Canada expressed the test of reasonable foreseeability in slightly more concrete terms:

80. *Ibid.*

81. *Ibid.*, at para. 14, citing *MacDonald v. Goertz*, 2009 BCCA 358 per Smith, J,A, at para. 17.

82. *Ibid.* See Linden and Feldthusen, *supra*, note 72 at 141.

83. *Rankin (Rankin's Garage & Sales) v. J.J.*, 2018 SCC 19 per Karakatsanis, J. at para 22, citing D. G. Owen, "Figuring Foreseeability" (2009), 44 *Wake Forest L. Rev.* 1277, at 1278.

84. *Rankin's (Rankin's Garage & Sales) v. J.J.*, 2018 SCC 19 at para. 23.

85. *Mustapha v. Culligan of Canada Ltd.*, *supra*, note 70 at para. 13.

"[w]hether or not an act or omission is negligent must be judged not by its consequences alone but also by considering whether a reasonable person should have anticipated that what happened might be *a natural result of that act or omission*."[86]

[Italics added]

Foreseeability of harm serves not only to determine to whom a duty of care is owed, but also to assess what that duty requires, and to delimit the scope of liability by considerations of remoteness. Even after breach of duty of care, causation and damage have all been proven, a court will not impose liability if the risk and the type of damage was too remote from the factual cause to have been reasonably foreseeable by the defendant.[87]

B. Artificial Intelligence and the Fault Principle

Artificial intelligence systems are rarely developed with the intent to do harm, but they may be intentionally used or adapted for harmful purposes. If humans or corporations intentionally use artificial intelligence to cause harm, fault can justly be attributed to those actors.

Application of the fault principle is more complicated when unintended harm results from the operation of an artificial intelligence system functioning with a high level of autonomy. Risk factors may be harder to predict and control than with other technologies, and when risks materialize, the source of the problem can be harder to identify. For example, was it the algorithm, the input data, or the way the system was operated? As autonomy increases, attribution of fault to particular humans or corporations becomes more tenuous.[88]

This is especially true with respect to systems that are based on machine learning. As noted earlier, these systems are highly dependent on the data used to train them. How a system that uses machine learning reaches a specific output may not be explainable.[89]

Data that a system may receive at any stage in its lifecycle may be inadequate, incorrect, or incomplete. How a system performed in training and testing with predefined

86. *University Hospital v. Lepine*, [1966] S.C.R. 561.

87. *Rankin (Rankin's Garage & Sales) v. J.J.*, *supra*, note 83 at para. 24, citing Linden and Feldthusen, *supra*, note 72 at 322.

88. Barfield and Pagallo, *supra*, note 16 at 16; Bathaee, *supra*, note 63 at 922.

89. Knight, *supra*, note 47.

datasets is not a certain guide to how it may react in actual use with new data that is not within the control of its designers and developers.

If no source of error can be identified to explain a harmful output, whether it be a bad decision, recommendation, or prediction, or the activation of a physical device, can the finger of blame be pointed at a particular defendant with any degree of certainty?

This context of autonomy, complexity, and data-drivenness will obscure and may make impossible to prove any causal link between an identifiable act or omission by a human and an output of an artificial intelligence system that causes harm.

It is also problematic to apply the concept of foreseeability where risk depends to a large extent on unknown variables introduced by future data inputs and unobservable digital processes that can adapt independently to improve their performance in pursuit of programmed goals. Bearing in mind that we are not speaking of behaviour that is found in the common store of previous human experience, what is a "real risk" of harm that should be foreseen and what is a "far-fetched" one such that it may reasonably be discounted without the potential for civil liability? What can be said to be a "natural result" of an act or omission by a human connected with the creation or deployment of the system, given the potentially infinite possibilities for intervening causal factors?

Autonomous artificial intelligence, especially when based on machine learning, challenges the application of concepts underlying the law of tort. Bathaee has said that "the law is built on legal doctrines focused on human conduct, which when applied to artificial intelligence, may not function."[90] In an article in the *Alberta Law Review*, two Canadian writers state:

> The practical challenges to the application of tort law principles to AI-related injuries are significant. Perhaps more daunting is the challenge Al presents to the very concept of fault that underlies much of Canadian tort law and advances its fundamental purposes. Given the nature of Al (and particularly machine learning aspects), an Al system itself may be beyond the effective context of designers, manufacturers, or users. This scenario is not amenable to current concepts of fault and causation in tort law.[91]

90. *Supra*, note 63 at 890-891.

91. Thomas O'Leary and Taylor Armfield, "Adapting to the Digital Transformation" (2020) 58:2 Alta. L.R. 249 at 262-263. Huberman, another Canadian writer, emphasizes that the designers and users of autonomously functioning AI systems do not have the level of control over risk that normally underpins liability in negligence: Pinchas Huberman, "Tort Law, Corrective Justice and the Problem of Autonomous-Machine-Caused Harm" (2021) 34:1 Can. J. Law & Jurisprudence 105 at 127.

Some may consider that view overstated, but the point is that when artificial intelligence is a factor, the application of conventional tort principles becomes considerably more difficult. The European Commission's High Level Expert Group on Liability and New Technologies has observed that the application of fault-based liability rules to emerging digital technologies is complicated by a lack of well-established models of their proper function and the fact that some of them develop by learning without direct human control.[92]

As explained below, the difficulty of applying conventional concepts of fault and foreseeability to the context of artificial intelligence has led some policymakers and legal theorists to advocate various special regimes of civil liability in which fault plays a lesser role, or no role at all. While that is not the approach taken in this report, an overview of different trends of thought on the matter of legal responsibility for artificial intelligence will provide perspective on our recommendations.

C. Competing Theories of Liability for Harm Caused by Artificial Intelligence

1. Strict Liability

Strict liability removes the need to prove fault on the part of a defendant. Only causation and damage need to be proven to establish liability. An early advocate for a rule of strict liability for harm produced by artificial intelligence was the U.S. writer David Vladeck, using autonomous vehicles as an example.[93] Vladeck and other proponents of strict liability acknowledge that it can be impossible to trace a failure by an autonomous device like a self-driving car to navigate correctly to a manufacturing or design defect, or another kind of human-caused fault like a programming error. They also take note that autonomous devices can be safer than human-controlled ones, so that their performance can be held to a higher standard. They proceed from this to the proposition that the concept of fault should be set aside entirely, and liability should be based instead on policy considerations.

92. Expert Group on Liability and New Technologies, New Technologies Formation, *Liability for Artificial Intelligence and other emerging digital technologies* (Brussels: European Union, 2019) at 23.

93. David C. Vladeck, "Machines without Principals: Liability Rules and Artificial Intelligence" (2014), 89 Wash. L. Rev. 117.

Vladeck contended that four policy reasons justified strict liability for harm caused by artificial intelligence. First, leaving innocent victims to bear loss runs counter to basic fairness, compensatory justice, and societal apportionment of risk, even if the cause is inexplicable. Second, the creators of an artificial intelligence system are in a position to either absorb costs of the harm or spread its burden widely through pricing. Costs of inexplicable damage should be borne by those who benefit from risk-reducing and innovative products. Third, strict liability spares everybody from "enormous transaction costs" of litigation over fault that cannot be established. Fourth, a predictable liability regime is more conducive to innovation than a less predictable one.[94]

Strict liability would have a place under a regime proposed in a 2020 resolution of the European Parliament proposing a draft regulation to the European Commission containing EU-wide rules on liability for harm caused by artificial intelligence systems.[95] The resolution was largely consistent with the 2019 report of the European Commission's High Level Expert Group on Liability and New Technologies.[96] It called for strict liability on the part of the operator of an artificial intelligence system classified as "high-risk" on the basis of a rather vague definition.[97] Liability of operators of systems that are not "high-risk" would be fault-based, with due diligence defences being available.[98]

The European Parliament resolution of 2020 provided that the liability of "producers" of artificial intelligence systems for harm caused by the systems should be

94. *Ibid.*, at 146-157.

95. European Parliament resolution of 20 October 2020 with recommendations to the Commission on a civil liability regime for artificial intelligence (2020/2014(INL).

96. *Supra*, note 92.

97. *Supra*, note 95, Annex, Art. 3(c). The definition of "high-risk" in the European Parliament resolution of 20 October 2020 is not the same as that in the EU *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts* of 21 April 2021, COM(2021) 206 final (the EU "*AI Act*"), or its later iterations generated in the EU legislative process. The grounds on which a system would be classified as high-risk under Article 6 and Annex III of the EU *AI Act* are much more specific.

98. *Supra*, note 95, Art. 8. The "due diligence" defence spelled out in para. 2 of Article 8 would consist of: "selecting a suitable AI-system for the right task and skills, placing it duly into operation, monitoring its activities and maintaining its operational reliability by regularly installing all available updates."

determined under the EU Product Liability Directive of 1985, which subjects "producers" to the regime of strict liability for defects in physical goods.[99]

At the same time, the European Parliament resolution urged revision of the Product Liability Directive "to adapt it to the digital world and to address the challenges posed by emerging digital technologies."[100] It stated the revision of the Product Liability Directive should include extension of the definition of "producer" to include manufacturers, developers, programmers, service providers, and "backend operators," the last-mentioned being defined as anyone "who, on a continuous basis, defines the features of the technology and provides data and an essential backend support service and therefore also exercises a degree of control over the risk connected with the operation and functioning of the AI-system."

The liability of operators of high-risk systems (though not that of producers) would be capped at two million euros for personal injury or death, and one million euros for verifiable economic loss or property damage.[101]

2. Product Liability

Some writers, chiefly in the U.S., have analogized cases of damage caused by artificial intelligence to those caused by a defect in a physical product, and have advocated for application of product liability principles.[102] These writers focus mainly on self-driving cars and robotic devices. They are viewing the landscape through the lens of U.S. product liability law, which in most states imposes strict liability for *manufacturing* defects, regardless of the degree of care taken by the manufacturer or others in the distribution chain to eliminate them or reduce risk.[103]

Liability for *design* defects under U.S. law turns on the application of one or the other of a "risk-utility" test or a "consumer expectations" test.[104] The risk-utility test asks whether the risk could have been reduced by an alternative design that could have

99. Council Directive of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products (85/374/EEC).

100. *Supra*, note 95, Preamble, para. 8.

101. *Supra*, note 95, Art. 5, para. 1.

102. F. Patrick Hubbard, "Sophisticated Robots: Balancing Liability, Regulation and Innovation" (2014) 66:5 Fla L Rev 1803 at 1852; Bryant Walker Smith, "Automated Driving and Product Liability" (2017) Michigan State L. Rev 1 at 15.

103. American Law Institute, *Restatement (Third) of Torts: Product Liability* § 1.

104. Andrew D. Selbst, "Negligence and AI's Human Users" (2020) 100 Boston University L. Rev. 1315 at 1323.

been adopted at a reasonable cost, and the omission to use the alternative design "renders the product not reasonably safe."[105] The somewhat older consumer expectations test determines whether the design makes the product dangerous beyond the extent that would reasonably be contemplated by a consumer.[106] These approaches resemble the analysis a court employs in a negligence case.

In contrast to the mainstream of U.S. product liability law, product liability claims made in Canada outside Québec are based on common law negligence, regardless of whether a claim is based on a manufacturing defect, a design defect, or failure to warn adequately of hazards known at the time a product is released into the market or subsequently discovered.[107] In the common law provinces and territories of Canada, proof of failure to exercise reasonable care by at least one defendant connected with the design, manufacture, or distribution of a product is essential in a product liability claim.

Québec is a civil law rather than a common law jurisdiction, and its laws regarding product liability are significantly different from those in the rest of Canada.[108] Fault is still an important element of a product liability claim in Québec, but a plaintiff may be able to take advantage of statutory warranties and presumptions to overcome difficulties of proof.[109]

There is a debate surrounding the question whether it is appropriate to characterize any form of artificial intelligence as a product, even if it is embodied in a physical

105. American Law Institute, *Restatement (Third) of Torts: Products Liability* (1998), §2(b).

106. American Law Institute, *Restatement (Second) of Torts* (1965), §402A.

107. The leading case grounding product liability in Canada in negligence is *Phillips v. Ford Motor Co. of Canada Ltd.* (1971), 18 D.L.R. (3d) 641 (Ont. C.A.). A few common law provinces have enacted statutory warranties that are binding on manufacturers of consumer products: e.g., *The Saskatchewan Consumer Products Warranties Act*, S.S. 1996, c. C-30.1*; Consumer Product Warranty and Liability Act*, S.N.B. 1978, c. C-18.1.

108. See Noah Boudreau and Nicolas-Karl Perrault, "What Lawyers, Manufacturers and Sellers Need to Know about Product Liability Laws in the Province of Québec" (Fasken Knowledge, 15 September 2020), online: https://www.fasken.com/en/knowledge/2020/09/15-what-you-need-to-know-about-product-liability-laws-quebec.

109. *Ibid.*

object like a robot or autonomous vehicle.[110] Some sources maintain artificial intelligence is better characterized as a service. Some sources maintain it is both.[111]

3. Notional Agency and Vicarious Liability

Another theory of liability advanced by some writers involves treating an artificial intelligence system as an agent, and the human or corporate deployer as the agent's principal.[112] Just as a principal is vicariously liable for the agent's acts and omissions within the scope of the authority given to the agent by the principal, the deployer would be liable for the results of the system's operation. The system's autonomy could be viewed like the authority a trusted agent could have to carry out tasks and achieve objectives set by the principal, but in a highly self-directed manner without micromanaged oversight.

A variant of the same idea is to treat the operator of the system like an employer of a human employee.[113] An employer is vicariously liable for acts and omissions of an employee while engaged in activities within the scope of employment.

Bucholz and Yu, two Canadian writers, have proposed "graded agency" as an appropriate liability regime for harm resulting from artificial intelligence.[114] This would be based on two principles, the first being that whoever adopts an artificial intelligence system for their own benefit should assume the risk of harm it may cause.[115] The second principle would be that the basis of civil liability for the risk should depend on the extent to which the defendant has delegated responsibility to the system.[116]

110. See Karni Chagal-Feferkorn, "Am I an Algorithm or a Product? When Products Liability Should apply to Algorithmic Decision-Makers" (2019), 30 Stanford L. and Policy Rev. 51.

111. See, e.g., http://www.differencebetween.net/technology/difference-between-ai-as-a-service-and-artificial-intelligence/.

112. The term "deployer" could extend to encompassing all or any of developers, suppliers, owners, users, operators, and anyone benefiting directly from operation of the system.

113. See Mihailis E. Diamantis, "Employed Algorithms: A Labor Model of Corporate Liability for AI" (2023) 72 Duke L.J. 797.

114. Ron Bucholz and Andy Yu, "Tort and Contracts: Civil liability for AI Causing Harm" in Jill Presser, Jesse Beatson and Gerald Chan, eds. *Litigating Artificial Intelligence* (Toronto: Emond, 2021) 348.

115. *Ibid.*, at 348.

116. *Ibid.*

If the defendant allows the system to operate entirely autonomously, as in using a fully autonomous vehicle, Bucholz and Yu say that strict (or vicarious) liability should apply for any harm the system causes.[117] If the defendant has maintained an ultimate decision-making role, this would be considered a partial delegation to the system and the defendant would be held to a negligence standard of reasonable care.[118]

4. Reasonableness of System Instead of Its Algorithms

A theory of liability advanced by Ryan Abbott and others holds that as autonomous artificial intelligence systems replace a human actor or decision-maker, tort law should treat them in the same way as a human decision-maker.[119] This would compel an analysis of whether the system actually functioned tortiously vis-a-vis the outside world, rather than attempting to trace the source of harm caused by the system to an error or fault on the part of the humans and corporations behind it. The theory is a corollary of Abbott's general premise of legal neutrality, meaning that law should make as few distinctions between human and artificial behaviour as possible.[120]

Proponents of this approach argue that in a negligence claim against a human tortfeasor, the thinking process that led to the tortious conduct is irrelevant vis-à-vis the victim. It is only the act or omission resulting in harm that matters, and whether it was due to a failure to meet the standard of care. They would maintain the same should hold in the case of a decision, act, or omission by an artificial intelligence system that would be tortious if the system was a human. In other words, the focus of the tort analysis should be on what the system actually did, rather than what made it act as it did.[121]

This approach would call for the acts or omissions of the system to be assessed on ordinary negligence principles, or in other words measured against the standard of care that would be applied to the conduct of a human actor.[122] The advantage for

117. *Ibid.*, at 349.

118. *Ibid.*

119. Abbott, *supra*, note 14 at 62-63.

120. *Ibid.*, at 3.

121. *Ibid.*, at 62-63.

122. Abbott notes that once artificial intelligence systems attain a level of safety and accuracy that is higher than what humans can attain, the standard of care applicable to them under this approach would be that of a system of the same kind, rather than the human standard: *supra*, note 14 at 9.

tort victims is that it would be unnecessary to prove that the system had an inherent defect. Defendants would have the opportunity to demonstrate that the system's outward act or omission was not negligent according to the same standard of care and degree of foreseeability that would apply to a human actor or decision-maker under the circumstances. If a human acting in accordance with the relevant standard of care could have caused the same accident in the same circumstances, there would be no liability because there was no negligence.

5. A Sliding Scale of Modified Liability Rules

Another approach that has been advanced calls for a sliding scale of liability regimes, depending on the level of transparency and autonomy of the system in question. Its proponents would apply current negligence principles where the deployment of the system is part of a human-driven decision-making process, such as a medical diagnostic system deployed to assist physicians who would make the final treatment decision.[123] Where a system acts autonomously, liability would turn on the degree of transparency (meaning the explainability of its processes), the constraints the creators or users placed on the system, and the extent of monitoring of the system in operation.[124]

In the case of autonomous operation, the creator or operator would be vicariously or strictly liable for the performance of the system when there is a high risk of harm. In less risky scenarios, originators and users would bear liability for negligence in testing, deploying, or operating the system. (Note this is akin to the dual regime of liability proposed by the 2020 European Parliament resolution and the scheme of "graded agency" proposed by the Canadian writers Bucholz and Yu.) If the system lacks transparency, a finding of negligence should depend on whether harm was a foreseeable consequence of deploying it to function autonomously, instead of whether the particular harm it caused was reasonably foreseeable.[125]

123. Yavar Bathaee, *supra,* note 63 at 894.

124. *Ibid.*, at 932.

125. *Ibid.*, at 938. See also Cynthia Khoo, "Missing the Unintended Forest Despite the Deliberately Planted Trees: Reasonable Foreseeability and Legal Recognition of Platform Algorithm-Facilitated Emergent Systemic Harm to Marginalized Communities." Draft paper presented at We Robot 2020 (3-4 April 2020, Ottawa, ON) at 73-74, online: https://drive.google.com/file/d/1xciaKKgI9QlsMM36ukK-LchG0dt0Rfws/view.

D. The Project Committee's View

1. Retention of Fault Principle vs. Strict Liability

The consultation paper contained a tentative recommendation that liability for damage caused by artificial intelligence should not be based on strict liability. The majority of responses to the consultation paper tended to support a rule of strict liability at least in relation to artificial intelligence systems used in or associated with high risk activities or contexts. Respondents favouring strict liability in relation to these systems offered various arguments for this position. One argument was the fact that a comprehensive regulatory regime for artificial intelligence is currently lacking. Another argument made was that negligence principles premised on what is reasonable to humans cannot be applied to machine learning, because the opacity of internal process prevents its outputs from being analyzed using reasonableness as a legal standard.

While acknowledging that not all artificial intelligence systems bring a high level of risk of harm, a leading research institute maintained in its response that some applications of artificial intelligence can properly be described as inherently dangerous, and for this reason it would not be a significant departure from Canadian tort principles to impose strict liability for harm resulting from them.

Things or activities that attract strict liability under the rule in *Rylands v. Fletcher*[126] are dangerous by nature and have a tendency to escape control. Some examples are fire, explosive substances, and wild animals.[127] Critics of strict liability would argue that comparisons between unpredictable results of machine learning and the escape of a wild animal or a noxious substance overrate the risk from artificial intelligence. They point to artificial intelligence systems that reduce or eliminate risk from human or machine activity in diverse settings, and would argue that autonomy of these systems that perform at a safer level than humans or technology operated directly by humans should be seen as a valuable quality rather than a source of danger.[128]

Ryan Abbott, a prominent critic of strict liability as applied to artificial intelligence, argues that excessive emphasis on risk associated with artificial intelligence ignores the greater safety and health benefits that it can bring where its capabilities exceed

126. See, *supra,* note 66.

127. Huberman, *supra*, note 91 at 130. Bucholz and Yu, *supra*, note 114 at 351-352 make an analogy between fully autonomous, "driverless" cars and wild animals, asserting both are inherently dangerous.

128. Huberman, *supra*, note 91 at 130.

those of humans. Treating artificial intelligence as an intrinsically dangerous thing, which is the effect of blanket adoption of strict liability, would foster a counter-productive liability and financial risk climate that would discourage development.[129] It is misguided to treat the deployment of artificial intelligence as tortious in itself.[130]

The Project Committee continues to agree with these criticisms of strict liability for the most part. An additional reason why the Project Committee is not in favour of strict liability, even in relation to systems at the higher end of the risk spectrum, is that it would not be as effective in encouraging the betterment of standards in the design, development, and use of artificial intelligence as fault-based liability. If firms know they will be held liable regardless of the degree of care they exercise, they will not have an incentive to raise and continually improve standards and practices that will increase safety. One respondent organization took issue with this position, maintaining that an incentive to limit risk exposure by reducing the frequency of harmful occurrences would remain even if developers and operators of artificial intelligence could not avoid liability in the event of one. While this might be true to an extent, we consider that the incentive to raise and maintain standards would not be as potent as under a fault-based regime where the prospect remains of avoiding liability altogether by meeting a legal standard of care.

Moreover, the implications of strict liability for the availability of insurance should also give pause to judges and legislators. If prospective insureds are held liable regardless of any degree of care taken to reduce risks that are not totally within their ability to eliminate, liability coverage is unlikely to be available at a manageable premium cost, or possibly not at all. Mainstream U.S. product liability law moved away from wholesale adoption of strict liability in the wake of an insurance crisis late in the 20th century, and now restricts it to manufacturing defects. Strict liability is not a guarantee that compensation will actually be recovered. It may only mean an easier path to a "dry" (unrecoverable) judgment. Withdrawal or contraction of liability coverage for the emerging technologies sector is not in the interests of plaintiffs.

The argument by opponents of strict liability that it would have a chilling effect on technological innovation is in part driven by a concern surrounding the availability of insurance.[131] This concern is especially acute in relation to smaller enterprises, where much original innovative work in digital technology takes place. An exodus

129. Abbott, *supra*, note 14 at 51.

130. Huberman, *supra*, note 91 at 130.

131. Karni A. Chagal-Feferkorn, "Am I an Algorithm or a Product? When Products liability Should Apply to Algorithmic Decision-Makers" (2019) 30 Stanford L. & Policy Rev. 61 at 81-82, 103 and 105.

of insurance providers from the AI sector could leave only the "big tech" giants that can self-insure active in the field.

Added to these reasons for backing away from strict liability even at what could be considered the high end of the risk spectrum is the difficulty of defining what "high-risk" or "high-impact" means in connection with artificial intelligence when there is as yet no clear consensus surrounding this. Legislators and policymakers are struggling with the task in the regulatory initiatives underway in various countries, including Canada. Creating different liability regimes based on a gradation of risk that is not well-established either in legislation or a demonstrated consensus of opinion would add further uncertainty to the law.

After giving consideration to all the arguments raised for and against strict liability, we remain convinced that a new rule of strict liability is unnecessary to deal with harm produced by artificial intelligence.

2. Notional Agency Insufficiently Distinguishable from Strict Liability

The theories based on notional agency and vicarious liability recognize that artificial intelligence systems are created to further objectives set by their human programmers and users. On a superficial level, they seem to benefit plaintiffs by avoiding the obstacles to proving fault on the part of the developers or operator that the opacity of the systems may present. The analogy to principal and agent is not a perfect one, however. It presents significant doctrinal issues.[132] An agent or employee has legal personality, and can be sued directly in place of or concurrently with the agent's principal or the employer for a tort committed while acting within the scope of the agent's authority or the employee's job. Artificial intelligence, however, lacks legal personality. Agents have fiduciary obligations of loyalty and good faith to their principals. It is hard to see how algorithms could fulfil these.[133]

In order to address some of the doctrinal issues, the suggestion has been raised that a new category of agent should be recognized to take account of the relationship between an artificial intelligence system and the humans who deploy it, that of "pure legal agent" without legal personality.[134]

132. Huberman, *supra*, note 91 at 143.

133. *Ibid.*

134. *Ibid.*

More practical questions are how to determine the scope of the system's "authority" and determine whether a given output that leads to harm is within or outside the "authority" of the system as a notional or "pure legal" agent. Principals and employers are not liable for agents and employees who engage in a "frolic of their own" outside the scope of their authority or employment and cause damage, e.g., an errant employee using a company vehicle recklessly as a party bus.[135] Without a workable answer to questions surrounding when emergent behaviour is to be considered within or outside the scope of the notional agent's authority, vicarious liability based on notional agency would differ little from strict liability in effect.

3. Non-Human Behaviour Is Not Measurable Against Human Reasonableness

The legal neutrality approach advocated by Abbott would call for the behaviour of an artificial intelligence system, rather than the conduct of the humans and corporations behind the system, to be measured against a standard of reasonableness. The advantage for tort victims is that it would be unnecessary to have to prove that the system or its data feedstock had an inherent defect or that something was done wrongly in training, testing, or operating it. Defendants would have the opportunity to demonstrate that the system's outward act or omission was not negligent according to the standard of care and degree of foreseeability that would apply to a human actor or decision-maker under the circumstances. If a human acting in accordance with the relevant standard of care could have made the same error in the same circumstances, there would be no liability because there was no negligence.

Arguably, this approach is fair to both plaintiffs and defendants. A plaintiff would face fewer obstacles in the path to recovery of compensation, and defences based on reasonable care would be available to all parties behind the system, e.g., an owner, operator, developer, or designer. This approach does not protect one at the expense of the other, and would not penalize innovation unduly, since it imposes no unusual liability rules. It has not gone unnoticed either that tort victims would be treated alike under this approach, regardless of whether the harm they incurred was caused by a human or by digital technology acting in place of one.[136]

The difficulty the Project Committee sees with legal neutrality as applied to the field of tort, however, is that reasonableness is a concept inextricably linked to the context of human behaviour and is defined by its appeal to human reason. Artificial

135. *Ibid.*, at 145.

136. Karni A Chagal-Feferkorn, "How Can I Tell If My Algorithm Was Reasonable? (2021), 21 Mich. Tech. L. Rev. 213 at 219.

intelligence and robots directed by it do not have human mental processes and a human's knowledge of the outside world. As a result, they err or fail in ways that are different from human error.

For example, in the case of the fatal collision between an autonomous test vehicle and a pedestrian walking a bicycle that was mentioned in Chapter 2, a human driver would not have failed to recognize the pedestrian as a pedestrian despite the presence of the bicycle. The navigation system of the test vehicle could recognize a pedestrian or a bicycle, but not the combined image.

As noted in Chapter 2, ChatGPT has been known to "hallucinate" fictitious citations when asked to generate a list of references.[137] It is reported to have composed a fictitious article and ascribed it to an actual author.[138] A human who was attempting to falsify citations would be extremely foolish to use the name of an actual author and would be unlikely to do so. In another example of difference between human and robotic error, GPT-3, a precursor of ChatGPT, is reported to have answered "yes" to the question "Is it safe to walk downstairs backwards if I close my eyes?" and "no" when asked a second time.[139] Asked the same question on another occasion, GPT-3 reportedly replied "That depends." While this could reflect learning behaviour, an adult human would likely answer that question consistently from the start.

We do not see reasonableness as a standard that can be applied coherently to harm caused by non-human actors. The tendency we foresee with the legal neutrality approach is for the non-human errors of artificial intelligence to be invariably categorized as unreasonable because a human would not make them in the same way, or because the task that the artificial intelligence performs is on such a scale that a human could not feasibly perform it in any event. This would result in a drift towards *de facto* strict liability until artificial intelligence equals or surpasses humans in terms of safety and reliability in all of its applications.[140]

137. *Supra*, note 58.

138. Geoff Brumfiel, "Here is What ChatGPT Gets Right – And Wrong" (NPR, 17 March 2023), online: https://www.npr.org/2023/03/17/1164383826/heres-what-the-latest-version-of-chatgpt-gets-right-and-wrong.

139. Gary Smith, "Chatbots: Still Dumb After All These Years" (MindMatters, 3 January 2022), online: https://mindmatters.ai/2022/01/will-chatbots-replace-the-art-of-human-conversation/.

140. Abbott, *supra*, note 14, notes at 65 that once the safety of artificial intelligence surpasses human standards, it would result in strict liability of humans if it became the standard of care.

If the performance of artificial intelligence is measured against a human standard of reasonableness, there is also a risk of reducing the incentive to exceed human capabilities in relation to safety and reliability.

4. Liability Upstream - Product Liability Provides Some Answers

The Project Committee considers that product liability law in the common law Canadian jurisdictions provides some answers regarding a just and balanced theory of liability while avoiding the difficulties of the other approaches discussed above. The principles concerning duty of care under product liability law may be applied by analogy to cases involving artificial intelligence without the need to take a definitive position on whether artificial intelligence itself may be characterized as a product.[141]

We would look to product liability law in regard to the liability of potential defendants situated "upstream" in the chain of events leading to litigation. Upstream defendants would be those involved in the design, development, training and testing of artificial intelligence systems prior to the point at which the systems are placed on the market or deployed in actual use. The liability of "downstream" defendants, namely operators and other end-users, involves different considerations and is discussed later.

A developer of a complete system that employs artificial intelligence may be realistically compared to a manufacturer of a complex product with numerous integrated components. It is common for digital technologies that employ artificial intelligence to incorporate discrete software modules created by different teams of specialists, much like a manufacturer of a complex physical product may draw on various sources for components.

The designers and developers of artificial intelligence modules that are incorporated into an integrated system by another developer may be compared to suppliers of components that are incorporated into a complex product. Suppliers of training and testing data in the developmental stage of a system may also be likened to component suppliers.

Product liability law recognizes a duty of care on the part of manufacturers and others who place a product in the stream of commerce towards anyone who may reasonably be foreseen to be at risk of damage or injury if the product is unsafe, not

141. Software has been characterized as a product subject to the *Sale of Goods Act* when it is integrated with hardware: *Burroughs Business Machines Ltd. v. Feed-Rite Mills (1962) Ltd.* (1973), 42 D.L.R. (3d) 303 (Man. C.A.); aff'd (1976), 64 D.L.R. (3d) 767 (S.C.C.).

only towards those who acquire the product or who deal in some manner with the manufacturer.[142]

The fact that courts have recognized this duty of care as existing in law relieves a plaintiff in a product liability case of having to specifically prove a relationship of proximity sufficient to support a duty of care owed to the plaintiff. If the relationship of proximity and the duty of care arising from it were not presumed in law, it would be impossible in most cases for a member of the public who is injured by a defective product and who never had direct or indirect dealings with the manufacturer to prove a proximate relationship. A member of the public who has incurred damage or loss through the operation of artificial intelligence will face the same obstacle in seeking justice if a duty of care similar to that of a manufacturer in product liability law were not extended by analogy to the developers of the system in question.

It appears to be the weight of opinion that component suppliers, as manufacturers in their own right, also owe a duty of care towards consumers of products containing their components and others who may be affected by defects in the components.[143]

The duty of care of a manufacturer of a complex product in which components made by others are integrated extends nevertheless over the entire product.[144] In other words, if a defect in a component makes the complex product dangerous, the manufacturer of the complex product is not exonerated from liability towards a plaintiff who suffers injury as a result of the defect merely because the supplier of the defective component is also liable.

142. *Bow Valley Husky (Bermuda) Ltd. v. Saint John Shipbuilding Ltd.*, [1997] 3 S.C.R. 1210, at para. 19. See also *Hollis v. Dow Corning Corp.*, [1995] 4 S.C.R. 634, at para. 21; *Stanway v. Wyeth Canada Inc.*, 2011 BCSC 1057; aff'd 2012 BCCA 260.

143. *Kett v. Mitsubishi Materials Corporation*, 2020 BCSC 1879, at para. 74. See also Theall, Lawrence G. et al., *Product Liability: Canadian Law and Practice* (Toronto: Thomson Reuters, 2000) (Looseleaf, updated) at 5-8. Despite some strong statements by text writers that component suppliers owe a duty of care to consumers and end users of a product containing their components, the direct liability of component suppliers to persons other than purchasers of their components does not seem to be as firmly established in case law as the duty of care of a manufacturer of the completed and integrated product. In *Burr v. Tecumseh Products of Canada Limited,*, 2023 ONCA 135, at paras. 103-104, however, the Ontario Court of Appeal expressly disapproved of an *obiter dictum* of the trial judge that a component supplier does not owe a duty of care directly to a consumer of a product in which the component is integrated.

144. *Winnipeg Condominium Corporation No. 36 v. Bird Construction Co.*, [1995] 1 S.C.R. 85; *1688782 Ontario Inc. v. Maple Leaf Foods Inc.*, 2020 SCC 35, at para. 49.

The principle that a manufacturer's duty of care extends over the entire product assists the plaintiff to gain access to civil justice more easily than if the plaintiff were limited to claiming against the supplier of the defective component alone, because it will generally be possible for a plaintiff to identify the manufacturer of a complex product marketed as a unit. It would often be far more difficult for a plaintiff to first determine the component in which the defect lay and then identify the source of the component.

The manufacturer of the complex product with integrated components is in a better position to identify its suppliers than the plaintiff, and will almost certainly assert a third party claim against the supplier of a defective component to bring the supplier into the litigation. If the manufacturer is sued, other proper defendants not likely to be as easily identified by the plaintiff are likely to be brought before the court.[145] The same would be true if a developer of a complex system employing an artificial intelligence system is sued and wants to shift blame to the suppliers of a software module the developer believes is likely connected with the plaintiff's claim.

The principles relating to duty of care under product liability thus provide a framework within the law of tort that can be adapted to allow rights and liabilities to be determined between someone who has been harmed by the operation of artificial intelligence and the upstream defendants who participated in developing the artificial intelligence system in question up to the point at which it is made available for actual use in real-world circumstances. Extending these principles by analogy to cases in which the operation of artificial intelligence has caused harm would facilitate access to civil justice by assisting plaintiffs to overcome the difficulty of identifying proper defendants at the upstream end and secure the appearance of other potentially liable defendants in the litigation.

Recommendations

1. Civil liability for harm caused by artificial intelligence should not be based on strict liability.

2. Product liability principles should be adapted by analogy to determine rights and liabilities as between a plaintiff harmed by the operation of an artificial intelligence system and defendants who participated in the development of the system and in making it available for use, by treating

145. A *proper* party is a person or entity with a legal interest in the matters in issue in a legal proceeding, and who therefore may join or be added to the proceeding as a party. By contrast, a *necessary* party is one who must be named in or be added to a legal proceeding as a party before the court can make a binding order resolving the matters in issue.

(a) the plaintiff similarly to a plaintiff claiming to have incurred loss or damage from a product comprising multiple components;

(b) developers of the system as owing a duty of care similar to that owed by a manufacturer of a complex product involving multiple integrated components towards persons or entities who foreseeably could be affected by a defect making the product dangerous;

(c) developers of components of the system as owing a duty of care similar to that owed by a supplier of a component of a complex product towards persons who foreseeably could be affected by a defect in the component that makes the component and the product in which it is integrated dangerous.

5. Liability Downstream - Operators

(a) Who is an "operator"?

In the previous section we used the term "upstream" to refer to the phase in the lifecycle of an artificial intelligence system covering the design, development, training and testing of an artificial intelligence system up to the time it is made available. We use the term "downstream" here to refer to the phase beginning with release of the system by its developers for actual use in real-world settings, and continuing for the life of the system.

In the downstream phase, users will deploy the system for their own purposes. They may have commissioned the development of the system on a custom-built basis, or they may have acquired it "off the shelf." They may lease the system from its developers. They may be the developers of the system themselves. They may operate the system directly or contract with an agent to operate the system on their behalf. The downstream users may be the developers themselves in some cases.

Regardless of how they come to have the system, or how they use it, users may expose others to the embedded risks it may bear or any risks that may emerge as the system continues to operate. Of course, the deployment of artificial intelligence in a user's activities may also improve their safety and have other positive effects. This does not change the fact that whoever has decision-making authority over the operation of the system is in a position to exert some level of control over risks associated with the system that may hold potential to affect others adversely.

The Project Committee believes that liability towards those affected by the system should go together with the ability to exert control over the risk of operation. This

aligns with the view expressed by the European Commission's High Level Expert Group on Liability and New Technologies in its 2019 report:

> [l]iability should lie with the person who is in control of the risk connected with the operation of emerging digital technologies and who benefits from their operation (operator).[146]

The ability to exercise control over the system risk in the course of operation should be sufficient in itself to attract a duty of care towards those whom the risk affects. We would not treat benefiting from the operation of the system as a requirement of the status of "operator." A requirement of benefit might be interpreted to exclude agents actively operating the system without having any beneficial interest in it from the status and liabilities of an "operator." There would be considerable moral hazard in allowing agents in active control to avoid liability towards those affected by the system merely because they do not own the system. The definition of "operator" should be broad enough to take account of many operating arrangements.

We would use the term "operator" to describe a person or entity with decision-making authority of a managerial nature over the operation of an artificial intelligence system and who thereby is in a position to exert some degree of control over the risk associated with its operation. We would make managerial authority a prerequisite for the status of "operator," because it would be unjust to impose the liabilities that should go with that status on technical personnel who carry out day-to-day system operation, but who are subject to superior orders and have no independent decision-making authority.[147]

146. *Supra*, note 92 at 39.

147. The consultation paper noted that this definition of an "operator" for purposes of tortious liability would correspond to language in the proposed Canadian federal *Artificial Intelligence and Data Act* ("AIDA"), *supra*, note 6, that would have imposed certain regulatory obligations in relation to an artificial intelligence system on a person who "manages" the operations of the system. The first reading bill contained a definition of "person responsible" for an artificial intelligence system that included a person who "manages its operation." The definition also included system designers, developers, and anyone making an artificial intelligence system available for use in the course of international or interprovincial trade and commerce. In November 2023, the Minister of Innovation, Science and Industry provided extensive draft amendments to the parliamentary Standing Committee on Industry and Technology reviewing the bill, which the government planned to introduce. At the time of writing of this report, Bill C-27 containing AIDA was still under review by the Standing Committee and had not yet been returned for third reading. The draft amendments would eliminate the definition of "person responsible," but would continue to impose obligations relating to safety and security of an artificial intelligence system on, *inter alia*, a "person who manages the operations" of a general-purpose system or a high-impact system, as defined elsewhere in the bill. This remains consistent with our view that the

The definition of "operator" we propose is flexible enough to allow for the possibility that two or more persons or corporate entities could have that status at any given time. For example, a system owner and the owner's agent operating the system under a contract could both be operators and have the same obligations vis-à-vis third parties whom the system may affect.[148]

A major legal organization responding to the consultation paper agreed with making managerial authority and control over risk the hallmarks of operator status, but noted there is a need to take account of situations in which one artificial intelligence system exercises decision-making authority over the operation of another. The respondent organization urged that the overseeing decision-making system be excluded from the definition of "operator" so as to preserve individual and corporate accountability in these situations. The Project Committee agrees that artificial intelligence systems should be excluded from operator status even if they do exercise oversight functions over other systems. The interposition of an overseer system between the human or corporate managers and a harm-causing system should make no difference with respect to the location of legal responsibility. Recommendation 3 below has been worded to take account of this type of situation.

(b) Liability of operators and other downstream defendants

As we have rejected strict liability, and envision problems with the application of other theories that have been advanced for reasons mentioned earlier in this chapter, the liability of operators and other potential downstream defendants such as providers of technical support services to the operator should flow in our view from the general fault-based principles that underlie the law of negligence. In other words, these downstream defendants should be held to owe a duty of care towards those within the foreseeable range of harm from the operation of an artificial

keystone requirement of the definition of an "operator" should be managerial authority over the operation of an artificial intelligence system.

148. The European Parliament's resolution of 20 October 2020 recommending a civil liability regime for artificial intelligence to the European Commission also contemplated that there could be more than one operator, but made a distinction between "frontend operators" and "backend operators." See, *supra*, note 95. This distinction, which was originally drawn in the 2019 report of the High Level Expert Group on Liability and New Technologies, *supra*, note 92 at 39 and 41-42, was meant to recognize which of two or more possible operators (such as an owner of an autonomous vehicle and the manufacturer who provides a continuous support service involving two-way transmission of data to and from the vehicle's systems) should bear strict liability because of having greater control over the risk associated with the technology. As the distinction is linked to a regime of strict liability, the Project Committee does not consider it necessary to recommend a similar distinction between categories of operators.

intelligence system or from the nature of their involvement with the system, and should be obliged to exercise reasonable care to prevent harm from arising.

The application of negligence principles in respect of both upstream and downstream defendants should be subject, however, to the modifications explained in the later chapters that we believe are required in the specific context of artificial intelligence.

(c) Recommendation

The Project Committee recommends:

3. (1) An individual or corporate entity with decision-making authority of a managerial nature over the operation of an artificial intelligence system and who thereby is in a position to exert some degree of control over the risk associated with its operation should be treated as an operator for the purpose of civil liability.

(2) A person or corporate entity described in paragraph (1) does not cease to be an operator merely because the operation of the artificial intelligence system in question is overseen or controlled by another artificial intelligence system.

4. The liability of operators and other persons who provide services in connection with the operation of an artificial intelligence system should be based on general principles of the law of negligence, subject to the recommendations made below.

E. Exclusion and Limitation of Liability for Artificial Intelligence

When digital technology is transferred or a licence is granted for its use, the risk of liability for damage caused by flaws in the technology itself or for damage that may arise from its use will typically be allocated by agreement. This usually means that the grantor of the rights will shift the risk to the party acquiring the rights by means of terms ("exclusion clauses") that prevent the other party from suing the grantor to recover the amount of a loss the other party may incur or the amount of a liability to a third party that is somehow related to the technology in question. Alternatively, liability between the parties may be limited by agreement to a fixed maximum.

Parties to a contract are generally able to agree to exclusion clauses, waivers of future claims, or limitation of damages, except to the extent that the ability to do so

may be restricted by law. Sometimes legislatures restrict the ability to contract out of liability or waive the right to make a claim for reasons of public policy. When legislatures restrict freedom of contract in this manner, it is usually to enforce specific legal or regulatory requirements, or to prevent erosion of certain rights conferred by law.[149] It is also possible for a contractual term to be found invalid for unconscionability,[150] or unenforceable because it contravenes an overriding public policy.[151] Apart from these circumstances, however, contractual terms excluding or limiting liability will usually be enforceable if they apply to the facts of a given situation.[152]

The EU initiatives described in this report prioritize protection of the public and enforcement of regulatory standards over the needs of developers and operators of artificial intelligence systems to manage risk exposure. The European Parliament resolution of 2020 on liability for harm caused by artificial intelligence referred to earlier in this chapter would render void any agreement between an operator and a person harmed by the operator's system that would limit or circumvent the obligations of the operator set out in the draft regulation attached to the resolution.[153] Similarly, the EU Product Liability Directive that would continue to govern damage claims against producers of artificial intelligence systems states that the liability of the producer under the Directive may not be limited or excluded by any contractual provision.[154]

---

149. Consumer protection legislation often contains anti-avoidance provisions that nullify waivers of rights and obligations under it. For example, s. 3 of the British Columbia *Business Practices and Consumer Protection Act*, S.B.C. 2004, c. 2, states: "3. Any waiver or release by a person of the person's rights, benefits or protections under this Act is void except to the extent that the waiver or release is expressly permitted by this Act."

150. *Tercon Contractors Ltd. v. British Columbia*, 2010 SCC 4, 2010 1 S.C.R. 69 at paras. 122-123 per Binnie, J. (dissenting in the result), with whom the majority agreed (at para. 62) regarding the appropriate legal framework for determining the validity of exclusionary contractual terms. See *Uber Technologies Inc. v. Heller*, 2020 SCC 16 (CanLII), [2020] 2 SCR 118 (majority finding arbitration clause in adhesion contract requiring arbitration in Netherlands unconscionable because it set up financial and logistical obstacles preventing the plaintiff from enforcing rights under the contract).

151. *Douez v. Facebook, Inc.* 2017 SCC 33, [2017] 1 S.C. R. 751 (choice of forum clause in online consumer adhesion contract unenforceable because of overriding public policy considerations, namely protection of Canadian consumers against massive inequality of bargaining power and preservation of the jurisdiction of local courts to interpret questions involving quasi-constitutional privacy rights).

152. *Tercon Contractors Ltd. v. British Columbia*, *supra*, note 150.

153. *Supra*, note 95, Art.2, para. 2.

154. European Union, *Council Directive of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products*,

Some might argue that the risks associated with autonomous artificial intelligence and emergence are such that restriction or removal of the ability to disclaim or contract out of liability for them is warranted. On the other hand, if developers and operators of artificial intelligence systems cannot manage their exposure to indeterminate liability as other businesses and research organizations must do, they may be unable to continue on an innovative path. This is especially true with respect to developers of generic or open-source artificial intelligence that can be used by any unascertained third parties and incorporated in applications of which the developer of the original generic or open-source module has no knowledge.

How an appropriate balance may be struck between adequate legal protection for the public and the need of industry and scientific organizations to manage their liability exposure in order to remain viable enterprises is a very complex question. Respondents to the consultation paper either did not comment on contractual exclusion and limitation of liability, or suggested only that it is a matter for legislatures to resolve on the basis of policy. Members of the Project Committee have divided views. As a result, this report does not contain a recommendation for any special rule of law concerning the extent to which it should be possible to disclaim, exclude, or limit liability for harm caused by artificial intelligence.

(85/374/EEC), Art. 12. The proposed new EU Product Liability Directive referred to in Chapter 4 would contain a similar provision.

Chapter 4. The Problem of Proof of Fault

A. General

In Chapter 3 attention was drawn to the likelihood that a plaintiff could face formidable problems of proof in relation to causation and fault in a tort claim arising from the operation of artificial intelligence. Writers have pointed to four principal reasons:

1. A high number of potential defendants typically involved in the design, development, deployment and operation of an artificial intelligence system.[155]

2. Autonomy of some systems;[156]

3. Limited explainability;[157]

4. Lack of foreseeability (as conventionally understood and applied).[158]

This listing of obstacles facing a plaintiff seeking compensation for damage is echoed by the European Commission in its proposal for an AI Liability Directive:

> Current national liability rules, in particular based on fault, are not suited to handling liability claims for damage caused by AI-enabled products and services…The specific characteristics of AI, including complexity, autonomy and opacity (the so-called "black box" effect), may make it difficult or prohibitively

155. Yaariv Benhamou and Justine Ferland, "Artificial Intelligence and Damage: Assessing liability and Calculating the Damages" in Pina D'Agostino, Carole Piovesan and Aviv Gaon, *Leading Legal Disruption: Artificial Intelligence and a Toolkit for Lawyers and the Law* (Toronto: Thomson Reuters Canada, 2020) at 170; High Level Expert Group on Liability and New Technologies, *supra*, note 92 at 28; European Commission, *Proposal for a Directive of the European parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)* (Brussels: European Commission, 28 September 2022) at para. (17) (preamble). See also Mihailis E. Diamantis, "Employed Algorithms: A Labor Model of Corporate Liability for AI" (2023) 72 Duke L.J. 797 at 808-809.

156. Benhamou, *supra*, note 155 at 170; Barfield and Pagallo, *supra*, note 88 at 16;

157. Benhamou, *supra*, note 155 at 170; Bathaee, *supra*, note 63 at 922.

158. Benhamou, *supra*, note 155 at 170; Bathaee, *supra*, note 63 at 924-925.

> expensive for victims to identify the liable person and prove the requirements for a successful liability claim.[159]

Pointing to a party at fault means first having to determine what actually caused a harmful output that produces damage. It is possible that only one component or system may be the source of damage, but there will often be multiple integrated systems running concurrently. The interdependency of software, hardware, and data in the functioning of artificial intelligence will tend to obscure the source of the harm.[160]

Even if a particular system can be identified as a source of harm, the limited extent of explainability that is associated with some forms of artificial intelligence, especially those that are highly data-dependent, will present a formidable obstacle to proving a claim. If the route from input to output is not fully explainable even by the designers and programmers of a system, plaintiffs have little hope of being able to present a detailed or step-by-step explanation linking a breach of duty by the defendant to the harm.

The difficulty a plaintiff would face in seeking evidence to prove causation and identify would be compounded by an inclination on the part of developers of an artificial intelligence system to treat its algorithms and other aspects of its design as proprietary secrets. This obstacle was illustrated in litigation concerning the dismissal of teachers by the Houston (Texas) Independent School District in reliance on a system marketed to school boards that used a proprietary statistical model for measuring every teacher's effectiveness. The supplier of the system withheld the algorithms and source code for the software that were in issue from both the plaintiffs and the defendant school district, preventing the plaintiffs from obtaining discovery of the basis for the automated evaluations and blocking independent testing of the algorithms that generated them.[161]

---

159. European Commission, *supra,* note 155, Explanatory Memorandum at 1.

160. Benhamou, *supra*, note 155 at 176;

161. *Houston Federation of Teachers, Local 2415 v. Houston Independent School District*, (4 May 2017) Civil Action H-14-1189 (U.S. District Court, So. Dist. of Texas), amended summary judgment opinion at 12. A subpoena was issued to the supplier of the system, who eventually gave an expert engaged by the plaintiffs very limited access to the source code: presentation by Martha P. Owen, counsel for the plaintiffs, at "AI Decision-Making: Protecting Rights Through Litigation and Regulation in Canada and the U.S.", web panel discussion sponsored by the Law Commission of Ontario, 9 December 2021. Regarding the issue of proprietary secrecy, see also Aidan Macnab, "School boards' lawyer suing social media platforms hopes trial reveals inner workings of algorithms," *Law Times*, 9 Apr. 2024.

Establishing the cause of the damage is essential to any tort claim, even ones based on strict liability.[162] Non-human, autonomous decision-making that is opaque places serious and potentially insuperable obstacles in the way of proving causation and fault.[163] These barriers to proof are likely to grow higher with increasing levels of autonomy and complexity As stated in an influential article in a U.S. law journal, artificial intelligence and robotics present society with "the prospect of a victim who suffers a non-natural harm but no perpetrator to whom the law can attribute this harm."[164]

B. Rebalancing Evidentiary Burdens for Fairness

1. General

Both upstream and downstream defendants will generally have a large advantage over the plaintiff in tort litigation involving artificial intelligence because of having greater familiarity with the system and knowledge of the measures that were or could have been taken to avert the harm that arose from its use.

Of course, the plaintiff can use oral and document discovery in the litigation process to elicit information helpful to the plaintiff's case, but the usefulness of pre-trial discovery as a means of redressing "informational asymmetry" between plaintiffs and defendants may be limited, particularly in cases involving data-dependent systems where the system's decision-making may be opaque due to the complexity of neural networks.[165] A discovery witness for a defendant developer or operator can be asked questions relating to facts, but cannot be asked on discovery to provide an opinion that would explain why certain input data produced an output resulting in damage. Information about the system derived from pre-trial discovery will not necessarily provide a basis for a theory of causation and fault that the plaintiff can plausibly advance.

162. *Ibid.*, at 171.

163. Benhamou, *supra*, note 155 at 175; Kristen Thomasen, "AI and Tort Law" in Florian Martin-Bariteau and Teresa Scassa, eds. *Artificial Intelligence and the Law in Canada* (Toronto: LexisNexis, 2022) at 117.

164. *Supra*, note 51 at 542.

165. The phrase "informational asymmetry" is used in the 2019 report of the European Commission's High Level Expert Group on Liability for New Technologies, *supra*, note 92 to describe the discrepancy between a plaintiff and the defendant operator and system developers with respect to knowledge of an artificial intelligence system in issue in a claim for compensation.

The extreme difficulties of proof that a plaintiff or prospective plaintiff may face in a claim for damages or other relief in a case involving artificial intelligence appears to warrant some mechanism to maintain balance in the litigation process in these cases. Numerous writers have reached this conclusion, as have EU policymakers.[166]

The concept of a rebalancing mechanism to achieve just results in the fact of an extreme informational imbalance, such as the application of evidentiary presumptions, is not entirely foreign to the law of tort. At an early stage in the development of the modern law of negligence, the common law developed the procedural device known as *res ipsa loquitur* to address cases where direct proof of causation and fault was lacking, but there was a strong likelihood of negligence.

Res ipsa loquitur ("the thing speaks for itself") was formerly applied in common law jurisdictions of Canada to allow for a rebuttable inference of negligence where three conditions were present: a thing or situation was under the sole control of the defendant or someone for whose actions the defendant was responsible, the occurrence resulting in damage would not happen in the ordinary course of things apart from negligence, and there was no evidence showing precisely how or why the damage occurred. If the defendant did not rebut the inference by providing an explanation of how the damage occurred that was consistent with the exercise of reasonable care, judgment would be given for the plaintiff despite the lack of clear proof of fault and causation. In effect, the plaintiff would benefit from a presumption of negligence drawn because of the lack of any other reasonable explanation for the occurrence resulting in damage.

In 1998, the Supreme Court of Canada held that *res ipsa loquitur* should no longer be treated as a distinct element of negligence law because of its "limited use."[167] The Supreme Court considered it to have been "more confusing than helpful," and declared it to have expired. Instead, the Supreme Court said courts should weigh circumstantial evidence together with direct evidence to determine whether the plaintiff has made out a *prima facie* (at first sight or impression) case of negligence against the defendant.[168]

The problem with requiring the plaintiff to make out a *prima facie* case of negligence in a case arising from harmful emergent behaviour of artificial intelligence is that it

166. See Benhamou and Ferland, *supra*, note 155 at 187; Barfield and Pagallo, *supra*, note 16 at 104. See also High Level Expert Group on Liability and New Technologies, *supra*, note 92 at 43 and 49-50.

167. *Fontaine v. British Columbia (Official Administrator)*, [1998]1 S.C.R. 424, at para. 27.

168. *Ibid.*

would require the plaintiff to present evidence on every element of negligence, including a specific act or omission amounting to fault on the part of one or more human or corporate defendants and a causal link between the fault and the damage when, at least in some circumstances, the nature of complex automated decision-making makes this impossible. This is especially true of data-dependent systems that employ deep learning.

This problem has been given considerable attention by EU policymakers. The proposal submitted to the European Parliament by the European Commission ("EC") on 28 September 2022 for an EU directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive Proposal) is chiefly concerned with creating a more equal playing field for claimants (plaintiffs) and defendants.[169]

The AI Liability Directive Proposal is one-half of the EC's response to the European Parliament's 2020 resolution mentioned in Chapter 3, in which the EC was invited to propose a special civil liability regime for artificial intelligence along the lines of the Annex to the resolution.[170] The other half of the EC's response package to the 2020 European Parliament resolution is its proposal for a new Product Liability Directive, which also contains provisions for claims based on the Directive that relate to artificial intelligence.[171] The next section describes the scheme under the EC proposals to rebalance the litigation process.

2. European Commission Rebalancing Proposals

(a) The AI Liability Directive Proposal

The AI Liability Directive proposal is intended to apply to fault-based claims under the national law of Member States. Its terminology and provisions are intended to mesh with those of the EU's proposed *AI Act*.[172]

The AI Liability Directive Proposal is less extensive than the regime of civil liability for artificial intelligence envisioned by the 2020 European Parliament resolution. Nevertheless, the mechanisms it contemplates would materially assist plaintiffs faced with obstacles to establishing fault and liability in certain circumstances.

169. European Commission, *supra*, note 155.

170. See the Explanatory Memorandum to the EC proposal for a new Product Liability Directive: European Commission, *Proposal for a Directive of the European Parliament and of the Council on liability for defective products* COM(2022) 495 final (Brussels: European Commission, 28 September 2022) at 6.

171. *Ibid.*

172. *Supra*, note 11.

The first assistive mechanism under the proposed AI Liability Directive would be an order to disclose relevant evidence concerning a specific "high-risk AI system," as defined under the proposed EU *AI Act*.[173] The proposed directive would require EU Member States to empower their national courts to order disclosure of relevant evidence about a high-risk artificial intelligence system suspected of having caused damage if a claimant has made "all proportionate attempts" to gather the evidence from the defendant.[174]

An application for a disclosure order could be made by a claimant (plaintiff) or a potential claimant who has requested the disclosure without success.[175] A potential claimant would have to present facts and evidence to support the plausibility of a damages claim.[176] A disclosure order could be made against a "provider" (developer) of an artificial intelligence system, a person subject under the proposed EU *AI Act* to the obligations of a provider, or a "user" (operator).[177] An order could be made against these persons whether or not they were defendants.

The disclosure order would be limited to evidence "necessary and proportionate" to support a claim or potential claim, taking the legitimate interests of all parties and third parties into account, including protection of trade secrets and confidential information.[178] The proposed directive would also oblige Member States to empower their courts to make orders for preservation of evidence that could be the subject of a disclosure order.[179]

Non-compliance by a defendant with a disclosure order or an order to preserve evidence would require a national court to presume that the defendant was non-compliant with a relevant duty of care to which the requested evidence relates and was

173. *Supra*, note 11, Art. 6, and its equivalents in subsequent versions of the *EI Act* generated in the EU legislative process.

174. European Commission, *supra*, note 155, Art. 3, para. 1.

175. *Ibid.*

176. *Ibid.* The language "evidence to support the plausibility of a damages claim" is drawn directly from the EC proposed directive. This wording points to an evidentiary standard that is likely similar to "establishing an arguable case" in common law jurisprudence.

177. *Ibid.*

178. *Ibid.*, Art. 3, para. 4.

179. *Ibid.*, Art.3, para. 3. The proposed directive and the accompanying explanatory text are ambiguous regarding whether an order for preservation could only be ancillary to a disclosure order, or whether a self-standing preservation order could be made without a corresponding order for disclosure.

intended to prove.[180] The defendant would have the right to rebut the presumption.[181]

The other mechanism in the European Commission proposal serving to relieve against the difficulties of proof in artificial intelligence-related claims for damages is a rebuttable presumption of a causal link applicable under certain circumstances. Courts of EU Member States would be required to presume a causal link between the defendant's fault and the output of the artificial intelligence system in issue, or its failure to produce an output, if:

(a) the claimant has demonstrated or the court has presumed (because of breach of a non-disclosure order) the fault of the defendant or a person for whose behaviour the defendant is responsible, consisting of non-compliance with a duty of care laid down in EU or national law directly intended to protect against the damage that occurred;

(b) it can be considered reasonably likely, based on the circumstances of the case, that the fault influenced the output produced by the artificial intelligence system or the failure of the system to produce an output; and

(c) the claimant has demonstrated that the output produced by the system, or the failure of the system to produce an output, gave rise to the damage.[182]

This presumption of a causal link would be rebuttable.[183]

The presumption would apply in a fault-based case involving any artificial intelligence system in an EU Member State, not only one classified as high-risk. There would be several important restrictions on its application, however. It is targeted principally at commercial or professional providers and users. If the defendant used the system in a personal and non-professional activity, the presumption would apply only if the defendant "materially interfered with the conditions of operation of the system," or if the defendant "was required and able to determine the conditions of operation of the system and failed to do so."[184]

180. *Ibid.*, Art. 3, para. 5.

181. *Ibid.*

182. *Ibid.*, Art. 4, para. 1.

183. *Ibid.*, para. 7.

184. *Ibid.*, Art 4, para. 6.

If the system is not classified as high-risk under EU law, the presumption would only apply where the national court considered it excessively difficult for the claimant to prove the causal link between the defendant's fault and the damage.[185]

In the case of claims against providers (developers) of high-risk systems, the requirement of paragraph (a) above would relate only to non-compliance with a specified set of statutory obligations applicable to those systems under the EU *AI Act*.[186]

In the case of defendant users (operators) of high-risk systems, the requirement of paragraph (a) would be met if the claimant proved the user did not comply with the obligation to use or monitor the system in accordance with the accompanying instructions of use or did not suspend or interrupt uses of the system where appropriate, as required by the *AI Act*.[187] It would also be met if the user exposed the system to input data under its control that was not relevant in view of the system's intended purpose, which would also be a breach of an obligation imposed on users by the *AI Act*.[188]

The presumption of a causal link would not apply if the defendant demonstrated that sufficient evidence and expertise was reasonably accessible to the claimant to prove the link.[189]

The AI Liability Directive Proposal is not intended to affect rights that a plaintiff would have under the laws of EU Member States implementing the product liability regime of the EU under the existing Product Liability Directive dating from 1985.[190]

185. *Ibid.*, Art. 4, para. 5.

186. *Ibid.*, Article 4, para. 2,

187. *Ibid.*, Art. 4, para. 3.

188. *Ibid.* Unlike Article 4, paragraph 2 concerning application of the presumption to providers, paragraph 3 does not say the presumption would apply against users *only* in the case of breach of these obligations. As a result, it is not clear from the language of Article 4, paragraph 3 whether the presumption of a causal link would be triggered under the proposed directive against a user of a high-risk system who is proven to have breached *other* obligations under EU or national law that intended to protect against the damage that is the subject of the claim.

189. *Ibid.*, Art. 4, para. 4. This appears circular, because in order to avoid having a causal link presumed under this exemption, the defendant would have to admit there is evidence already available to the plaintiff to prove the causal link between the fault of the defendant and the damage, thereby admitting the causal link.

190. *Ibid.*, Article 1, para. 3(b).

That directive imposes a regime of strict liability for defective products.[191] Most claims in the EU against developers of artificial intelligence systems would likely be made under the Product Liability Directive because it avoids the need to prove fault.

(b) The EC proposal for a new Product Liability Directive

The EC's proposal for a new Product Liability Directive is intended to "ensure liability rules reflect the nature and risks of products in the digital age."[192] It would repeal the 1985 directive, but carry forward the regime of strict liability of manufacturers for damage from defective products.[193] It would also clarify that artificial intelligence systems are "products" within the scope of the new directive, settling doubts on whether they come within the EU product liability regime.[194] This will likely result in most future claims in the EU for damages against developers of artificial intelligence systems being made under national laws implementing the Product Liability Directive rather than under laws requiring proof of fault.

Even under a strict liability regime, however, the plaintiff must still prove a causal link between a defect in the product and the damage. The proposed new Product Liability Directive contains provisions to alleviate difficulties of proof in product liability claims by means of presumptions once a reduced evidentiary threshold is met. It would also require Member States to empower their courts to order a defendant in a claim under the Product Liability Directive to disclose relevant evidence requested by a claimant on a basis very similar to the terms of the disclosure order provisions of the AI Liability Directive.[195]

A product would be presumed defective if the claimant established that the product does not comply with mandatory safety requirements under EU or national law within Member States that are intended to protect against the damage that occurred, or if the damage that is the gist of the claim was caused by an obvious malfunction

191. European Union, *Council Directive of 25 July 1985 on the approximation of the laws, regulations and administrative provisions of the Member States concerning liability for defective products*, (85/374/EEC), Art. 1.

192. *Supra*, note 170, Explanatory Memorandum, p. 2. In March 2024, the European Parliament resolved to approve the proposed new Product Liability Directive with some amendments, allowing it to move forward to the next stage of approval by the Council of the EU.

193. *Ibid.*, Art. 7, para. 1.

194. *Ibid.*, Explanatory Memorandum, p. 5. "Product" is defined in Art. 4(1) to include software. The debate elsewhere regarding whether artificial intelligence is a product, a service, or a hybrid product and service is noted in Chapter 3 under the subheading "Product Liability."

195. *Ibid.*, Art. 8.

during normal use or under ordinary circumstances.[196] It would also be presumed if the defendant did not comply with an order to disclose relevant evidence.[197]

A causal link between defectiveness of the product and the damage would be presumed if the defect is established and the damage is of a kind typically consistent with the defect in question.[198] If the court determined that the claimant faced excessive difficulties in proving defectiveness of the product or a causal link with the damage because of technical or scientific complexity, both defectiveness and the causal link between it and the damage would be presumed if the claimant demonstrated that "it is likely that the product was defective or that its defectiveness is a likely cause of the damage, or both."[199]

The presumptions to alleviate difficulties of proof under the new Product Liability Directive would all be rebuttable, and the defendant could contest the existence of excessive difficulties of proof facing the plaintiff.[200]

3. Recommendation on Relief Against Difficulties of Proof in Appropriate Cases

The solutions to the evidentiary problems that are proposed in the EU are highly statutory and prescriptive in nature. It may not be necessary to go the length to which the European Commission has gone in recommending that courts be obliged by legislation to draw inferences of fault and causation in order to address the practical difficulties of proof that will face plaintiffs seeking compensation for harm caused by artificial intelligence. The European Commission's approach appears inflexible, depriving courts of the discretion to respond to the circumstances of individual cases.

The solution we will recommend instead is one that need not be legislated, but could equally as well be applied by courts as the need arises on the facts of individual cases. Not every case involving artificial intelligence will present problems of proof of fault and causation. Reversing the ultimate onus of proof resting on the plaintiff in a civil action only because artificial intelligence is involved in the factual matrix

196. *Ibid.*, Art. 9, para. 1.

197. *Ibid.*

198. *Ibid.*, Art. 9, para. 3.

199. *Ibid.*, Art. 9, para. 4.

200. *Ibid.*, Art 9, para. 5.

would be an extreme measure, coming close to imposing a regime of technology-specific strict liability.

The European Commission proposals also depend on a distinction between systems classified as "high-risk" and others. This is a distinction that relates to the choice to impose an extensive body of specific regulatory standards at the Union level on systems legislatively classified as "high-risk," and to the dual regimes of strict and fault-based liability under the proposed EU schemes that have been discussed above. As we have rejected strict liability in principle, we have not adopted the concept of gradations of risk in our recommendations on civil liability for artificial intelligence. While classifications of systems based on risk level may be appropriate in relation to differential regulatory burdens, we do not see it as necessary or desirable to distinguish between risk levels in relation to evidence and procedural matters.[201]

In the consultation paper, we made two tentative recommendations (numbered 5 and 6) to deal with proof of causation and fault in claims for damages arising from the operation of artificial intelligence. Tentative Recommendation 5 would have applied at the pre-trial discovery stage of a tort action. It was based on the product liability analogy between the developer of an integrated system incorporating artificial intelligence and the manufacturer of a complex, multi-component product. It would have imposed an onus on a defendant who made the complete, integrated system available for use to provide an explanation for its performance in the circumstances of the case that was consistent with the exercise of reasonable care in its design, development, training, and testing, or else disclose sufficient information concerning the design and function of the system to other parties (including co-defendants) to allow the other parties to advance a theory of causation.

Tentative Recommendation 6 in the consultation paper would have allowed the court to infer negligence against the defendants jointly and severally in the event the developer of the complete system failed to discharge the onus under, except as

201. A prominent research and policy organization responding to the consultation paper took issue with this stance, arguing that the legal framework surrounding AI use should correspond to the level of risk even in procedural matters, because the great diversity of contexts in which AI is deployed makes risk a more appropriate distinction to reflect in law than the fact of AI use. The respondent also considered that strict liability should attach to high-risk uses of AI. Introducing procedural distinctions on the basis of risk gradations would add unnecessary complexity to the law, however, and would overlook the fact that data-intensive and statistical AI applications are capable by their nature of presenting obstacles to proof of causation and fault in tort claims, regardless of the degree of risk and scope of potential harm present in the use cases where the claims arise.

against a defendant who rebutted the inference by producing evidence of reasonable care on the part of that defendant.

Respondents to the consultation paper criticized these tentative recommendations as impractical, potentially requiring disclosure that was not technically possible or that would be resisted because they could require breach of contractual obligations of confidentiality owed to parties higher in the supply chain or to third parties outside it. Objection was also raised to Tentative Recommendation 5 on the ground that it appeared to allow the plaintiff, rather than the court, to determine whether disclosure was sufficient in a given case.

A further criticism raised against Tentative Recommendation 5 in the responses was that imposing an onus to produce an explanation for the performance of the artificial intelligence component on a defendant positioned downstream from the designers and developers of that component was no less unfair than placing the onus on the injured plaintiff, since the internal processes of the component may be equally opaque to both the plaintiff and downstream defendants.

The Project Committee continues to view difficulty in gaining pre-trial discovery of the details of algorithms and the training and testing of artificial systems as a serious issue that parties in litigation relating to artificial intelligence will face, whether they be plaintiffs or defendants. The Houston Independent School District case mentioned at the beginning of this chapter is highly illustrative of the problem.[202] Tentative Recommendations 5 and 6 were designed to induce developers and operators to provide discovery of crucially relevant information out of self-interest to avoid triggering an inference of negligence in the absence of an explanation of the facts consistent with lack of negligence.

The confidentiality of commercially sensitive technical information relevant to matters in issue is routinely protected in litigation through implied undertakings of confidentiality by all parties, express undertakings given to the court, and orders for sealing of records and transcripts, all enforceable by sanctions for contempt.[203] In this regard, artificial intelligence need not be treated differently from other

202. See *supra*, note 161 and accompanying text.

203. Information obtained through oral and documentary discovery in litigation is subject to an implied undertaking by the examining party to whom it is disclosed to use the information only for purposes of the litigation in which it is obtained, unless the scope of the implied undertaking is varied by order of the court: *Juman v. Doucette*, 2008 SCC 8, [2008] 1 S.C.R. 157 at para. 4. In addition, orders for sealing documentary exhibits and transcripts in order to protect trade secrets and confidential information included in the evidence are commonly made on application by a party.

technologies that lead to litigation. The Project Committee also continues to consider the principle drawn from product liability that the duty of care of a manufacturer of a multi-component product extends to the entire product as one that is appropriate to extend to the developer of an integrated digital system incorporating artificial intelligence.

Nevertheless, we have heard the criticisms respondents raised concerning the combination of Tentative Recommendations 5 and 6 as set out in the consultation paper. In this report we have drawn back from recommending the introduction of any special discovery obligations at the pre-trial stage that would be peculiar to cases involving artificial intelligence.

We continue to believe, however, that courts must apply legal principles of causation and proof with awareness of the nature of artificial intelligence technology when plaintiffs incur damage or loss from it and are unable, because of the opacity of artificial intelligence decision-making, to enunciate a provable theory of why the system generated the harm-producing output in the particular circumstances, or identify specific acts or omissions in the design, development, training, testing, or operation of the system that are causally linked to the damage incurred. To insist that a plaintiff must prove causation and fault with that degree of exactitude without sufficient regard to the limited explainability of some forms of artificial intelligence could amount to acceptance of "the prospect of a victim who suffers a non-natural harm but no perpetrator to whom the law can attribute this harm."[204]

Once a plaintiff in a case of this kind has proven that the output of an artificial intelligence system has produced loss or damage, defendants who participated in the design, development, training, testing, and deployment of the system should expect to be called upon to provide evidence that in their respective roles, they exercised reasonable care to prevent or reduce the likelihood of the system occasioning harm. The relevant facts surrounding the genesis and deployment of the system are within the collective knowledge of the defendants. If they are unable to show that they exercised reasonable care, it is not unjust for them to bear responsibility for the harmful output.

The mechanism we propose to relieve against the obstacles to proof that the nature of artificial intelligence may present in some cases is a rebuttable inference of fault and causation that would come into play at the conclusion of trial after consideration of all the evidence. It could be drawn if reasonable care has not been proven and no explanation for the behaviour of the system arises from the evidence that is consistent with the exercise of reasonable care to prevent harm of the kind incurred

204. Calo, *supra*, note 51 at 542.

by the plaintiff.[205] The inference would not be drawn, of course, against defendants who are found to have exercised reasonable care in the role they individually played in the continuum from initial design to eventual deployment and use of the artificial intelligence system.

A rebuttable inference of this kind is consistent with the principle recognized by the Supreme Court in *Clements v. Clements*[206] that liability may be found on the part of multiple defendants when, through no fault of the plaintiff, it is impossible to prove on the balance of probabilities which acts or omissions of specific defendants in breach of a duty of care actually caused the harm, but it is clear that some or all of the defendants have materially contributed to the risk of harm occurring and each of them can point the finger at the other. As the majority in the Supreme Court observed in *Clements*, to allow defendants to avoid liability in these circumstances would be "at odds with the fairness, deterrence, and corrective justice objectives of the law of negligence."[207]

The Project Committee recommends:

5. Except as against any defendant who is found to have exercised reasonable care in the circumstances leading to an action for damages or other relief due to harm to persons or property arising from the operation of artificial intelligence, a court deciding such an action should be justified in drawing an inference that a lack of reasonable care on the part of defendants responsible for the design, development, training, testing, or use of the system is causally linked to the harm incurred by the plaintiff, if

(a) the harm alleged by the plaintiff is proven to have been caused by the output of the artificial intelligence system, either functioning alone or as a component of an integrated system;

(b) the evidence taken as a whole does not establish the exercise of reasonable care by defendants in the design, development, training, testing, and use of that system or yield an explanation for the behaviour of the system in the circumstances of the

205. While this rebuttable inference would somewhat resemble the adverse inference (sometimes referred to as a shift in the evidentiary burden of proof) that formerly could be drawn under the obsolete maxim *res ipsa loquitur*, it is also consistent with the approach approved by the Supreme Court in *Fontaine v. British Columbia (Official Administrator)*, *supra*, note 167 of weighing circumstantial evidence and drawing appropriate inferences of fault in the absence of evidence to contradict the *prima facie* case of negligence.

206. *Supra*, note 73.

207. *Ibid.*, at para. 32.

case that is consistent with the exercise of reasonable care by those defendants; and

(c) due to the characteristics of the artificial intelligence system, the plaintiff cannot reasonably be expected to identify specific acts or omissions by specific defendants that caused or materially contributed to causing the system to occasion the harm.

C. Reasonable Foreseeability and Artificial Intelligence

The test of foreseeability is currently expressed as what a reasonable person in the position of the defendant would consider a "real risk" that is "not far-fetched" or a "natural result" of the defendant's act or omission.[208] While it is described in law as an objective standard because it is based on the imputed perceptions of a reasonable person, it is still one that is rooted in human experience and human perceptions of cause and effect and risk materialization. The reactions of a non-human, artificial decision-making process to an infinite quantity of potential inputs are outside that experience.

Risk variables associated with artificial intelligence include inability to control the quality of input data in the future use of a system by third parties ("garbage in, garbage out"), the unpredictability associated with extreme complexity, autonomous operation, and emergence. The variables influencing the risk associated with artificial intelligence, especially machine learning, will tend to confound the concept of reasonable foreseeability as it is conventionally understood and applied in the law of negligence.[209] The American writer Bathaee has commented on this subject:

> [t]he result of the AI's decision or conduct may not have been in any way foreseeable by the AI's creator or user. For example, the Al may reach a counter-intuitive solution, find an obscure pattern hidden deep in petabytes of data, engage in conduct in which a human being could not have engaged (e.g., at faster speeds), or make decisions based on higher-dimensional relationships between variables that no human can visualize. Put simply, if even the creator of the AI cannot foresee its effects, a reasonable person cannot either. Indeed, if the creator of AI cannot necessarily foresee how the AI will make decisions, what

208. See Chapter 3 under the subheading "Unintended Harm."

209. Thomasen, *supra*, note 163 at 113;

> conduct it will engage in, or the nature of the patterns it will find in data, what can be said about the reasonable person in such a situation?[210]

Another risk variable that cannot be discounted, although it may possibly be more easily weighed under the current test of reasonable foreseeability because it relates to human behaviour, is deliberate and inventive misuse of the artificial intelligence by third parties.[211]

In a leading case on foreseeability in Canadian tort law, *Rankin (Rankin's Garage and Sales) v. J.J.*, the Supreme Court of Canada held that the risk of personal injury through the reckless driving of a juvenile car thief was not reasonably foreseeable by a garage owner who had stored the car unlocked on the garage lot with the keys in an ashtray.[212] The car thief's behaviour pattern after the theft in *Rankin* could be seen as comparable to the emergent market-gaming activities of a stock-trading algorithm in Bathaee's hypothetical example mentioned in Chapter 2.[213] Should the owner of the algorithm in that hypothetical be held liable to investors who lost money due to its manipulation of the market? The Supreme Court's answer would seem to be "No."

In an actual example of autonomous, emergent behaviour that was probably even less likely to have been foreseen by creators, a popular chatbot was reported to have produced a fictitious news report naming an actual individual as the offending party in an entirely fictitious incident of sexual harassment.[214] Furthermore, it appeared that after the erroneous content was suppressed on the chatbot platform where it originated, a counterpart chatbot repeated the defamatory material.[215]

---

210. *Supra*, note 63 at 924.

211. Intentional misuse is an especially acute problem in relation to online platforms and other openly available online services that can be easily adapted to malevolent purposes such as spreading misinformation for political or ideological purposes or creating internal discord in a targeted society or community. See Khoo, *supra*, note 125. Canadian courts have held manufacturers liable for not taking cost-effective measures to protect against relatively obvious possibilities of misuse; see *Tabrizi v. Whallon Machine Inc.* (1996), 29 C.C.L.T. (2d) 176 (B.C.S.C.). Reckless misuse of a product in disregard of instructions has been held not to be foreseeable, however: *Lem* v. *Barotto Sports Ltd.*, [1976] A.J. No. 4412(S.C., App. Div.).

212. *Supra*, note 83.

213. *Supra*, note 63.

214. Pranshu Verma and Will Oremus, "ChatGPT invented a sexual harassment scandal and named a real law prof as the accused" (*Washington Post*, 5 April 2023), online: https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/

215. *Ibid.*

Reasonable foreseeability can be retained as a "crucial limiting principle"[216] of liability in negligence cases involving artificial intelligence as in other cases, but operators and developers should not be heard to say that the risk of harmful emergence altogether (as opposed to a specific occurrence of emergence) is unforeseeable. If this were the case, victims of harm related to the operation of artificial intelligence would be less protected under tort law than victims incurring similar injury or loss at the hands of human tortfeasors.

Asymmetric protection of this kind should be avoided as a matter of legal policy, as it would detract from two of the principal goals of tort law, compensation and prevention of harm.[217] If the threshold of tortious liability is set lower for defendants responsible for artificial intelligence than for defendants in other tort cases, it will not encourage adequate risk assessments and maintenance of best practices in the field.

The concept of reasonable foreseeability will need to be applied differently than in the general run of negligence cases where artificial intelligence is not a factor. Foreseeability of risk in relation to the use of artificial intelligence must be considered not only in terms of particular outcomes, but in terms of the risk of unpredictability itself. Rather than be applied with regard to the particular manner in which risk materialized in the given case, reasonable foreseeability must be understood as treating unpredictability and emergence as known risks that potentially give rise to unknown ones.[218]

In order to continue to act as a limiting principle preventing indeterminate liability, however, the determination of what was reasonably foreseeable should be made with reference to the known attributes of the system in question at the relevant time, the use cases for which the system was intended, and known or predictable

---

216. *Rankin (Rankin's Garage & Sales) v. J.J.*, *supra*, note 83 at para. 23.

217. The potential for asymmetric protection of the law as between victims of artificial intelligence and victims of human tortfeasors was a significant concern of EU policymakers, reflected in one of the governing principles listed in the resolution of the European Parliament of 20 October 2020 for a proposed civil liability regime for artificial intelligence, *supra*, note 95:

> 7 Citizens should be entitled to the same level of protection and rights, irrespective of whether the harm is caused by an AI-system or not, or if it takes place physically or virtually, so that their confidence in the new technology is strengthened.

The asymmetric protection issue is addressed at some length in the report of the European Commission's High Level Expert Group, *supra*, note 92.

218. Thomasen, *supra*, note 163. See also Expert Group on Liability and New Technologies, *supra*, note 92 at 45.

alternate uses (including predictable misuse) of the system. In our view, these considerations delineate the scope of potential harm that designers, developers, and operators should be expected to consider in assessing risk and taking preventive measures.

A prominent research organization responding to the consultation paper raised the question whether the relevant time for assessing known attributes of the system should be taken as of the occurrence of harm or the time of use in the particular case in question, rather than at the time when the system was initially deployed or released for real-world use. The respondent suggested that treating the attributes of the system as known at the later time (e.g., occurrence of harm) as the criterion would serve the policy goal of encouraging continuous monitoring of system performance over its lifecycle.

The Project Committee considered this point at length. It was acknowledged that making continuous monitoring of a system's performance an element of the standard of care is a worthy policy objective, and in fact Chapter 5 urges that Canadian courts should treat it as such. It is nevertheless a troubling question whether a fair result is possible if the determination respecting the foreseeability of a risk had to be made with reference to attributes of a system that became known only at a later time than the creation of the system and the deployment.

Designers and developers of an artificial intelligence system that suddenly displays harmful emergent behaviour only after being placed in actual use are in a somewhat similar position to a manufacturer of a product containing a latent hazard that results in harm to consumers after it has been released on the market. The manufacturer would be liable to the consumers for the damage if the hazard could have been discovered by reasonable means before the product entered the market. Once the hazard has come to light, the manufacturer has a duty of care to warn users of the risk that is now known. The manufacturer is not held liable on the basis of what could not have been known or discovered beforehand. Ultimately, the Project Committee concluded that the usual temporal frame of reference for assessing reasonable foreseeability need not shift or be modified only because damage resulted from the operation of artificial intelligence.

The Project Committee has not attempted to re-formulate the test of reasonable foreseeability in a semantic manner. Instead, our recommendation calls for the application of the test to be modified in cases involving harm causally connected with artificial intelligence in order to recognize the potential for "known unknowns"[219]

219. See Thomasen, *supra*, note 163 at 118.

when a decision is made to deploy artificial intelligence or to make it available for deployment.

The Project Committee recommends:

6. Where harm results from the operation of an artificial intelligence system and a claim based on negligence is made, the test of reasonable foreseeability of harm should be applied with regard to the risk that the system might behave unpredictably to cause harm in an unknown manner, taking into account

(a) attributes of the system known at the relevant time;

(b) intended use of the system; and

(c) known or predictable alternate uses of the system.

Chapter 5. Standard of Care

A. How the Standard of Care Is Set

In order to determine whether a duty of care has been met or breached in a negligence claim, the court compares the defendant's conduct against the standard of care. The court must make a determination regarding what the content of the standard of care is in the circumstances of the case. This usually involves an assessment of the degree of care a reasonable person in the defendant's position would exercise. In making that determination, the court can draw upon evidence given in the case before it, including expert evidence, about what the accepted or customary standard is in the industry or profession in question in the action. It can draw upon decisions regarding the standard of care in previous similar cases. The court can also look to relevant legal and regulatory requirements affecting the activity in issue in the case.

The standard of care is a common law concept, and is sometimes set on the basis of policy. Legislation and regulations governing the activity in issue are relevant considerations in setting the standard of care in an individual case, but not decisive. Breach of a statute or regulation applicable to the defendant or the defendant's activities in issue does not automatically lead to a finding of liability. The Supreme Court of Canada has explained the relationship between regulatory enactments and the standard of care as follows:

> Legislative standards are relevant to the common law standard of care, but the two are not necessarily co-extensive. The fact that a statute prescribes or prohibits certain activities may constitute evidence of reasonable conduct in a given situation, but it does not extinguish the underlying obligation of reasonableness. See *The Queen (Canada) v. Saskatchewan Wheat Pool*, [1983] 1 S.C.R. 205. Thus, a statutory breach does not automatically give rise to civil liability; it is merely some evidence of negligence. See, e.g., *Stewart v. Pettie*, [1995] 1 S.C.R. 131, at para. 36, and *Saskatchewan Wheat Pool*, at p. 225. By the same token, mere compliance with a statute does not, in and of itself, preclude a finding of civil liability. See…[citation omitted]. Statutory standards can, however, be highly relevant to the assessment of reasonable conduct in a particular case, and in fact may render reasonable an act or omission which would otherwise appear to be negligent. This allows courts to consider the legislative framework in which people and companies must operate, while at the same time recognizing that one cannot avoid the underlying obligation of reasonable care simply by discharging statutory duties.[220]

220. *Ryan v. Victoria (City)*, [1997] 1 S.C.R. 201, at para. 29.

B. A Largely Open Playing Field – For the Time Being

As yet, there is little regulation of the field of artificial intelligence in Canada. That may change under the proposed *Canadian Artificial Intelligence and Data Act* ("AIDA")[221] which is before Parliament at the time of writing. As introduced, AIDA took the form of a framework statute that would leave much detail to future regulations and would only apply to some artificial intelligence systems. It would require a a "person responsible" for an artificial intelligence system to assess whether it is a "high-impact system" as defined in regulations. If so, the person responsible would then be required to establish various measures in accordance with the regulations to identify, assess, and mitigate risks of harm or biased output that could result from the use of the system.[222]

Shortly after the bill containing AIDA was introduced, the Government of Canada issued a Companion Document explaining the policy behind the legislation and listing factors that could be applied to determine whether a system would qualify as "high-impact."[223] The requirements outlined in the AIDA Companion Document for high-impact systems align generally with the themes found in recent policy framework documents for artificial intelligence issued by other countries: human oversight and monitoring, transparency, fairness and equity, safety, accountability, validity (performing consistently with intended objectives) and robustness (stability and resilience in a variety of circumstances.)[224]

AIDA is likely to be heavily amended before it is enacted. In November 2023, the Minister of Innovation, Science and Industry provided the text of amendments the

221. *Supra*, note 6.

222. AIDA, supra, note 6, s. 8. Proposed amendments would eliminate the definition of "person responsible," but continue to impose risk identification, assessment and mitigation obligations and maintain the overall regulatory scheme of AIDA. See note 147, *supra*.

223. The *Artificial Intelligence and Data Act (AIDA) Companion Document* issued by the federal Department of Innovation, Science and Economic Development provides some indication of what future regulations might contain, but also emphasizes that the development of regulations will occur after a 6-month consultation on their content. It also states AIDA will not be in force until two years after Bill C-27 is passed, and at the earliest no sooner than 2025. See online: https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document..

224. *Ibid.*

Government intended to introduce later to the parliamentary Standing Committee considering the bill.[225] The amendments include a new definition of "artificial intelligence system" that resembles the OECD definition reproduced in Chapter 2. This new definition would expand the scope of AIDA considerably because it would no longer be confined to systems that process data related to human activities. A definition of "high-impact system" would be placed in the Act itself instead of a regulation.[226] The amendments would also address multi-purpose artificial intelligence systems as a distinct category under a definition of "general-purpose system."[227]

The Government of Canada *Directive on Automated Decision-Making*[228] applies to artificial intelligence systems used by federal bodies that make recommendations or decisions about individuals outside government. The Directive contains risk reduction and management provisions that have value as models that could be transferred to other public and private sector settings.

In September 2023, the Canadian government announced a *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems*, which it described as being intended to foster responsible development and operation of "advanced generative systems with general-purpose capabilities" and mitigate their risks in the interests of public safety "in advance of binding regulation" AIDA.[229] Firms adhering to the Voluntary Code undertake to implement a list of measures that include: a comprehensive risk management framework taking into account risks associated with inappropriate or malicious uses, "red-team" (adversarial) testing to identify vulnerabilities, curating training datasets to manage data quality and detect biases, making guidance on appropriate system usage available to downstream developers and managers, human oversight and monitoring, and

225. Letter from the Hon. François-Philippe Champagne, Minister of Innovation, Science and Industry to Joël Lightbound, MP, Chair, Standing Committee on Industry and Technology, undated, accompanying text of proposed motions for amendment of Bill C-27, clause 39.

226. *Ibid.*

227. *Ibid.*, s. 5(1) as proposed, definition of "general-purpose system."

228. *Supra*, note 6 .

229. Innovation, Science and Economic Development Canada, *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems* (27 September 2023), online: https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems. Among the major AI developers and research centres that have adhered to the Voluntary Code are Blackberry, Telus, IBM, Cohere, OpenText, Vector Institute, Mila, and Alberta Machine Intelligence Institute (AMII): https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems#wb-auto-4.

clearly identifying AI systems as automated when their interactions could be mistaken for human ones.[230]

The relatively few regulatory standards for artificial intelligence in place at the present time and lack of precedents in case law will mean that Canadian courts will be called upon to break new ground in setting the standard of care in the early cases that come forward. Some guidance can be drawn from international regulatory precedents that have begun to appear, such as the proposed EU *AI Act,*[231] U.S. presidential *Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence* of 30 October 2023,[232] the proposed U.S. *Algorithmic Accountability Act*,[233] public policy framework documents such as the *White House Blueprint for an AI Bill of Rights*,[234] NIST *Artificial Intelligence Risk Management Framework*,[235] the UK Government policy paper *A pro-innovation approach to AI regulation*,[236] and published voluntary guidelines issued by various scientific and industry bodies. These sources may have considerable advisory and persuasive value pending the development of a body of Canadian precedent, and possibly afterwards as well.

230. *Ibid.*

231. *Supra*, note 11, and as later amended in the EU legislative process.

232. EO 14110 of Oct. 30 2023, 88 FR 75191, online: https://www.govinfo.gov/content/pkg/FR-2023-11-01/pdf/2023-24283.pdf.

233. H.R 6580, 117th Congress, 2nd Sess.

234. See online: https://www.whitehouse.gov/wp-content/uploads/2022/10/Blueprint-for-an-AI-Bill-of-Rights.pdf.

235. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1 (Washington: Dept. of Commerce, January 2023).

236. Department for Science, Innovation and Technology (UK), *A pro-innovation approach to AI regulation*, Policy paper (29 March 2023), online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf. In February 2024, the UK Government issued its response to the consultation that followed issuance of the March 2023 Policy Paper, reiterating its intention to rely on a pro-innovation policy combining five cross-sectoral principles to be applied by regulators in their context-specific mandates (1. safety, security and robustness 2. appropriate transparency and explainability 3. fairness, 4. accountability and governance 5. contestability and redress) that were outlined in the March 2023 paper as well as voluntary measures by developers rather than on statutory regulation. The February 2024 response holds the door open for later legislative initiatives to mitigate potential AI-related harm, especially in relation to "highly advanced general purpose models": Department of Science, Innovation and Technology, *A pro-innovation approach to AI regulation: government response*, Cmnd CP 1019 (London: Department of Science, Innovation and Technology, 6 February 2024), online: https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response at paragraphs 65-77.

Courts will usually look to accepted practice within an industry or profession in making a determination about the standard of care in cases involving that industry or profession. Industry and practice standards are relevant to the standard of care in negligence in the same way as regulatory enactments; they have persuasive value, but are not decisive.[237] A court may make a finding of negligence against a defendant who has conformed to an industry practice if it finds the practice creates an unreasonable risk and the standard of care should be set higher.[238]

While our recommendations on the basis of liability and proof of fault do not distinguish between systems on the basis of risk level for reasons given earlier, the level of risk associated with a system undoubtedly will be among the factors weighed by courts in determining the standard of care applicable in a given case.[239] In so doing, courts should not lose sight of a reasonable balance between risk and benefit, and the need to avoid discouraging innovation by imposing an excessively onerous burden of liability.

Artificial intelligence, like other digital technology operating through the internet, is not limited by territorial boundaries in its reach and its effects. Despite the fact that there are very few hard and fast rules governing the field of artificial intelligence at the present time, an interjurisdictional consensus on basic elements of good practice in the development and use of artificial intelligence is emerging. This can be seen in the similarities between the contents of policy framework documents in various countries and in relatively consistent ethical and technical guidelines generally recognized as amounting to norms of good practice in the research community and software industry even though, as noted in a response to the consultation paper, current actual practice may often fall well short of these guidelines.

The convergence of norms of good practice is occurring despite the marked differences between jurisdictions in approach to regulation of artificial intelligence. The strongly prescriptive regulatory approach taken in the EU contrasts with the announced policy of the UK Government to minimize new regulation in the interests of promoting innovation.[240] At the same time, the UK policy paper calls for existing sectoral regulators in the country to apply principles broadly similar to those

237. *Ding v. Prévost*, 2022 BCSC 215 at para. 211; *Zsoldos v. CPR*, 2009 ONCA 55 at para. 43; leave to appeal to S.C.C. refused 33083 (9 July 2009).

238. *Zsoldos v. CPR*, 2009 ONCA 55 at para. 43.

239. See text under the subheading "3. Recommendation on Relief Against Difficulties of Proof in Appropriate Cases" in Chapter 4.

240. *Supra*, note 236.

embraced elsewhere, and notes the importance of international alignment and "interoperability with international regulatory frameworks."[241]

As the interjurisdictional consensus is evolving and is still in a relatively early stage of development, we will not make a recommendation about the substance of the standard of care in a negligence case involving artificial intelligence. Instead, our recommendation urges that courts set the standard of care in litigation concerning artificial intelligence with a worldwide outlook in order to take account of interjurisdictionally recognized best practices and the desirability of encouraging responsible innovation.

The balance of this chapter lists some of the elements of good practice in the design, development, and operation of artificial intelligence systems that appear with relative consistency in an international cross-section of regulatory policy documents and instruments, and thus appear to enjoy wide recognition. We think Canadian courts would do well to consider them when called upon to set a standard of care in civil litigation arising from the operation of artificial intelligence.

C. Widely Recognized Elements of Good Practice

1. The Design, Development, Training, and Testing Phases

(a) Transparency

Transparency regarding what a system is intended to do, how the system does it, who brought it into being, and who is using it is considered of paramount importance. At a minimum, there should be disclosure of:

- the identities of the producer, owner, and user of the system;
- capabilities and limitations of the system;
- accuracy levels for which the system has been validated, the metrics used to validate the accuracy levels, and other characteristics of performance;
- known risks from intended use and foreseeable patterns of misuse;
- where the system has a machine – human interface, the fact that the human is interacting with an automated system;

241. *Ibid.*, at 43.

- notification to persons affected by an automated decision that the decision is being made by automation.[242]

Transparency regarding the risks and limitations of an artificial intelligence system is especially crucial if the system is released on an open-source basis that can be obtained and used by innumerable parties, a point that was emphasized both in the deliberations of the Project Committee and by respondents to the consultation paper.

Generic systems (also known variously as "general-purpose AI" or "foundation models") have been a particular focus of current regulatory initiatives because of concern over the level of risk that may arise from their use by third parties in applications of which the developer of the generic system has no knowledge. The proposed amendments to AIDA referred to earlier would impose special transparency requirements regarding "general-purpose systems," defined to mean artificial intelligence systems that are designed for use or adaptation in many fields or for many purposes and activities, including ones not contemplated during the development of the system.[243] The most recent text of the proposed EU *AI Act* also contains special rules for providers of "general-purpose AI models" that include documentation requirements and provision of required information and documentation concerning the model and

242. In Quebec, anyone whose personal information is used by an enterprise to render a decision based exclusively on automated processing of that information has the right after 22 September 2023 to be informed of the personal information used to render the decision, the reasons, principal factors, and parameters leading to the decision, and of the right to have the personal information corrected: s. 12.1 of *An act respecting the protection of personal information in the private sector*, CQLR c. P-39.1, s. 12.1, as am. by S.Q. 2021, c. 25, s. 110. The Information and Privacy Commissioner, the Ombudsperson for British Columbia and the Yukon Ombudsman and Information and Privacy Commissioner have recommended in a joint report that public authorities and service providers be required to notify individuals when an automated decision system is being used to make decisions about them, and to explain the operation of the system in an understandable fashion: Office of the Information and Privacy Commissioner and Ombudsperson (B.C. and Yukon): *Getting Ahead of the Curve: meeting the challenges to privacy and fairness arising from the use of artificial intelligence in the public sector*, Joint Special Report No. 2 (Victoria and Whitehorse: June 2021), Recommendation 2 at 47. The joint report also recommends that privacy legislation be amended to provide a right to: notification that an automated decision system is in use, explanation of the criteria used by it, and an ability to object to the use of the automated decision system: Recommendation 7(b) at 48.

243. *Supra*, note 225, proposed motion 039-087-34a_EN for amendment of s. 5(1) of AIDA to add definition of "general-purpose system." The requirements for general-purpose systems would include a plain-language description of: the system's capabilities and limitations, and risks of harm or biased output from any reasonably foreseeable use of the system additional information to be prescribed by regulation. They would also include any prescribed measures to ensure that the public can identify any text, image, audio, or video output as having been generated by artificial intelligence: proposed new text of s. 7(1)(f).

the content used for training it to parties downstream who intend to integrate the model into another system.[244]

Thorough documentation of the system's features, function, and risks, prepared in conformity with transparency requirements, is essential. The system documentation should be available to all personnel operating the system interpreting its output. The documentation should explain how a system was developed and tested so as to inform downstream developers, operators and users of the uses for which it has been validated. For example, it is essential for prospective users of a medical artificial intelligence model to be given the details of the population from which data was derived to develop the algorithms, as it may be unsuitable for use on a population with different age, genetic background, or medical history characteristics.

At least early in the lifecycle of an AI system, if not throughout it, the system documentation should include information on in-use monitoring that reflects the risk factor in the use of the system.

(b) Vital design features

Systems should be designed to achieve consistent levels of accuracy, robustness, and cybersecurity appropriate for their intended uses.

The possibility of feedback loops whereby biased outputs become inputs in later operation of systems that continue to learn while in use after deployment is something that needs to be mitigated.

Of course, foreseeable risks should be taken into account in the design of a system and efforts made to minimize the possibility of their arising.[245] This should include taking measures where possible to limit the scope of harm that could arise if a known risk materialized, a point made by a leading research organization in its response to the consultation paper.

244. *European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts* (COM(2021)0206-C9-0146/2021 – 2021/0105(COD)), P9_TA(2023)0138, Article 53 and Annex XII, online: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.pdf.

245. See the discussion under the subheading "C. Reasonable Foreseeability and Artificial Intelligence" in Chapter 4.

(c) Data Quality and Data Governance

Good data governance and management practices require in particular:

- the splitting of training, validation and testing datasets;
- ensuring relevance, completeness, robustness of data;[246]
- measures to eliminate bias in input data;
- statistically rigorous data collection conforming to legal requirements;
- compliance with standard practices in data cleaning;[247]
- documentation of the data governance process.

(d) Continuous risk assessment

Continuous risk assessment throughout a system's lifecycle, covering, *inter alia*:

- risk not only arising in the course of intended use, but also from foreseeable misuse;
- cybersecurity risks, including adversarial exploitation and data poisoning (malicious attack on the system through adulterating or manipulating input data);
- the possibility of bias in output;

246. "Robustness of data" as used here means that the data is accurate and representative on the average with respect to the population to which it relates. The term "robustness" is used in more than one sense in connection with artificial intelligence. It is also used to denote the ability of an artificial intelligence system to generate accurate and reliable output despite the presence of some inaccurate or irrelevant data (so-called "noise") within a dataset, or the ability of a system to perform well under varying parameters of use. In another sense, "robustness" also refers to the resilience of a system to cyberattack.

247. Data cleaning or "scrubbing" is "the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset": Tableau, online: https://www.tableau.com/learn/articles/what-is-data-cleaning.

- consideration of various risk mitigation measures referred to variously as: "outlier detection," "anomaly detection," "data sanitation," and employing ones recognized as effective and appropriate;

- validation that large-scale statistics on deployment resemble those from small-scale test sets;

- assessment and mitigation measures for risk to children and other vulnerable populations.

(e) Independent validation

Pre-deployment review of a new system by at least one expert who has not been involved in its design and development is seen as desirable or imperative, depending on the jurisdiction and intended use of the system.

Whether engagement of a knowledgeable at-arm's-length third party to review or independently validate a system should be regarded as part of the legal standard of care should depend on the risk associated with use of the system and the risk factors in the particular use case.

(f) Articulation of appropriate human involvement

The design of an artificial intelligence system needs to include an articulation of where a human will be involved, what decisions the human will make, and what data the human will use in making them.

(g) Logging Capability

The capability of a system to create an automatic record of its own operation, including features of transparency, is of high importance.

(h) Monitoring

Monitoring of system performance by the developer and operators throughout the system's lifecycle.

Performance monitoring by a third party at arm's length from the developers and suppliers of a system with sufficient expertise to evaluate system performance is also necessary in areas of high risk, including the medical domain.

Effective monitoring requires a continuing flow of information on system performance between users, developers, and any third party engaged in monitoring the system.

Monitoring of system performance is not to be equated with constant human oversight. Making constant human oversight a standard would negate much of the benefit of using artificial intelligence. Further, much research and actual events have shown it is unwise to rely exclusively or even primarily on human oversight as a safety mechanism. Evidence suggests that humans are not especially adept at deciding when they should assume direct control of an automated system, and the design and operational characteristics of a system may hinder them in reacting appropriately when they re-assert control.[248] The U.S. National Transportation Safety Board put it in these terms: "When it comes to the human capacity to monitor an automation system for its failures, research findings are consistent – humans are very poor at this task."[249]

(i) Updating

Updating and upgrading to correct problems as they arise from monitoring of the system while deployed in actual use.

2. Operation in Actual Use

(a) Risk Assessment and Mitigation

The operator needs to carry out its own risk assessment to complement that of the system developers, as the operator designates and controls the specific setting in which the system will be used.

Regulatory and governance models emphasize that risk assessment needs to be an iterative process continuing throughout the lifecycle of the system.

The risk assessment should comprise what the developer's risk assessment would cover and at least the following in addition:

248. Elish has warned against the legal and moral pitfalls of misappropriating responsibility for technological failures to "the nearest human operator" of an automated system, despite the limited control the human actor may have had over the system: Madeleine Clare Elish, "Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction" (2019) 5 ESTS Journal 40, online: https://estsjournal.org/index.php/ests/article/view/260/177.

249. National Transportation Safety Board, *Collision Between Vehicle Controlled by Developmental Automated Driving System and Pedestrian: Tempe, Arizona March 18, 2018* NTSP/HAR 19/03 PB 2-19-101402 (Washington: NTSB, 2019) at 44.

- cybersecurity risks specific to the operator's organization and other activities;
- data privacy risks and mitigation measures in the event of a breach;
- impacts on affected populations, including those that do not consist of direct users of the operator's system.

(b) Compliance with Developer's Recommendations and Terms of Service

Compliance with the recommendations of the developer for operating the system within its design parameters is a basic element of good practice. Off-label use of a system would presumably heighten risk.

If the operator is involved in the development or modification of a system, it should fall to the operator to provide all necessary information in its possession or power to clarify the use case for the benefit of designers and developers.

(c) Transparency

Persons interacting with the system or who are directly affected by its output should be made aware by the operator in clear, readily understandable terms that they are dealing with an automated system, are informed of its purpose, and have its output explained to them.[250]

If a system generates or manipulates visual or audible content to resemble existing persons, objects, places, or representations of events and could deceive someone into thinking the content was authentic, original, or accurate, the operator must at least have made it known or obvious that the content has been artificially generated or manipulated ("deepfaked"). It must also be open to a court to find artificial generation or manipulation to be a breach of the standard of care in itself, depending on

250. Section 12.1 of Québec's *Act respecting the protection of personal information in the private sector*, CQLR c. P-39.1, as enacted by S.Q. 2021, c. 25, s. 110 requires that organizations provide notice to consumers of an automated decision using their personal information at the time the decision is made. It confers several additional related rights on consumers, including the right to be informed of the reasons, principal factors, and parameters taken into account in the decision, and to correct the personal information used to make it. In Ontario, an amendment to the *Employment Standards Act, 2000*, S.O. 2000, c. 41 will require employers who use artificial intelligence to screen, assess, or select applicants for a publicly advertised job to disclose its use, subject to exemptions by regulation: see the *Working for Workers Four Act, 2023*, Bill 149, 1st Sess., 43rd Legislature, Sch. 2, s. 2(1), adding s. 8.4 to the to the *Employment Standards Act, 2000*, S.O. 2000, c. 41. The amending Act had passed third reading as of the date of this report.

the circumstances. An example would be deepfaking done for deceptive, malevolent, or abusive purposes.

Legally authorized or permissible security and surveillance systems and systems used in law enforcement to detect or investigate criminal conduct may be exceptions to basic transparency requirements, but may sometimes require substantially higher standards of care.

(d) Monitoring of System Performance

Continuous monitoring of performance is needed throughout the life of the system, including co-operating with the developer's post-deployment monitoring process and supplying performance data for that process as necessary. Monitoring with respect to risk mitigation measures will become a statutory obligation for operators of high-impact systems (as they may come to be defined) under the proposed Canadian *Artificial Intelligence and Data Act*.[251]

System performance monitoring should be both a human and automated process where circumstances warrant, so that the human and automated monitors act as a check and verification of each other. The performance of human personnel responsible for system oversight should be included in the overall monitoring process.

Metrics must be created to measure the system's performance. The metrics themselves should be subject to periodic evaluation. A range within which the system is intended to operate must be specified, so as to enable monitoring for model drift.

Scaling issues (reliability of the system under expanded loads and parameters) also need to be covered in the monitoring process.

Data quality must be continually monitored for bias and to ensure the current training data remains relevant to the intended purpose of the system.

Re-use of data in new contexts requires extra care to prevent spreading and amplification of harms that the original data had the potential to produce. If outputs are used as inputs in subsequent use of a system, such as may occur with systems that learn continuously in operation, data that is derived or inferred from prior outputs of the same system should be treated as high-risk for feedback loops that could compound and amplify bias or inaccuracy in the previous outputs.

---

251. *Supra*, note 6. The relevant provision of AIDA in the first reading version of Bill C-27 was s. 9. The proposed amendments to the bill referred to, *supra*, in note 147 would create additional monitoring requirements and redistribute them in several sections of AIDA..

(e) System Maintenance

Appropriate preventive and mitigative maintenance could include:

- updating the system as recommended by the developer, or to correct problems identified through continuous monitoring;
- re-training the system or modifying the system model in response to issues detected, including model drift and change in external conditions;[252]
- retaining a baseline version of the system model to compare results of later updated, re-trained, or modified versions;
- reverting to an earlier baseline system that worked satisfactorily if reconfigurations and retrained models prove unreliable.[253]

(f) Privacy

Compliance with applicable laws on data privacy, as well as adherence to sector-specific ethical standards.

(g) Training

Appropriate training for personnel responsible for the operation of the system.

(h) Logs and Other Record-keeping

Creation and retention of logs in keeping with any regulatory requirements and protocols established by the operator or recommended by the developer of the system.

Retention of monitoring results, along with records of any corrective steps taken, so that the history of any problems with the system and mitigative efforts can be tracked.

The proposed Canadian *Artificial Intelligence and Data Act* would require that records be kept relating to the establishment of data anonymization, risk assessment,

252. WH Blueprint, *supra*, note 234 at 19.

253. *Ibid.*

risk mitigation, and monitoring protocols.[254] Additional record-keeping requirements may be added by amendments before passage of the Act or prescribed by regulation afterwards.

(i) Organizational governance framework for artificial intelligence systems

The operator's organization should have clear governance structures and procedures for oversight and risk mitigation. Responsibilities of designated individuals or groups for these functions should be clearly defined.

Responsibility for decisions concerning problems and potential shutdown or "roll-back" from any change should rest at a level in the operator's organization that will allow fast response, with anyone holding this decision-making authority being fully informed with respect to risk potential.

D. Recommendation

The Project Committee recommends:

7. (1) The standard of care in litigation concerning artificial intelligence should be set so as to encourage responsible innovation and development, encompassing reasonable care to avoid foreseeable injury or loss.

(2) In setting the standard of care, courts should employ a broad interjurisdictional perspective, extending where applicable to

(a) nationally and internationally recognized best practices;[255]

(b) national and international regulatory standards;

in the design, development, and operation of artificial intelligence systems.

---

254. *Supra*, note 6. The reference is to s. 10(1) of AIDA in the first reading version of Bill C-27, Part 3. Proposed amendments referred to, *supra*, in note 147 would add further record-keeping requirements.

255. Including the examples listed in Section C of this chapter, but this recommendation is not intended to detract from the ability of a court to find that a prevailing practice or custom in an industry, profession, or community falls short of a reasonable standard of care, regardless of the extent to which it is observed. See *Waldick v. Malcolm*, [1991] 2 S.C.R. 456 at 473; *Roberge v. Bolduc*, [1991] 1 S.C.R. 374 at 436-437; *Zsoldos v. CPR*, *supra*, note 237.

Chapter 6. Algorithmic Discrimination and Civil Liability

A. Bias: A Recognized Problem in Artificial Intelligence

The potential for artificial intelligence to produce biased outputs with discriminatory effects is widely recognized as being among the most significant legal and ethical problems associated with this technology. Virtually all public policy and regulatory initiatives across the jurisdictions emphasize the necessity of eliminating, averting, or mitigating biased output.[256] At the international level, member states of the Council of Europe are in the process of negotiating a treaty that is largely aimed at ensuring the development and expansion of artificial intelligence stays in keeping with human rights and freedom from discrimination.[257]

The term "bias" is used in at least two senses in connection with artificial intelligence outputs.[258] One is technical and is drawn from the field of statistics: a systematic error that renders a body of data incorrect on average with respect to the

256. See, for example, the proposed Canadian *Artificial Intelligence and Data Act*, *supra*, note 6, ss. 5 (definition of 'biased output"), 8 and 9 (in first reading version of Bill C-27, Part 3); Government of Canada *Directive on Automated Decision-Making*, *supra* note 6, para. 6.3.1; European Union proposed *AI Act*, *supra*, note 11, as adopted by the European Parliament on 13 March 2024, art. 10, paras. 2(f) and (g), and art. 15, para. 4; White House *Blueprint for an AI Bill of Rights*, *supra*, note 234 at 23-29; UK Dept. for Science, Innovation & Technology, *A pro-innovation approach to AI regulation*, Cmnd 815 (London: 29 March 2023), online: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1146542/a_pro-innovation_approach_to_AI_regulation.pdf.

257. See Council of Europe, Committee on Artificial Intelligence, *Consolidated Working Draft of the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law* (7 July 2023), online: https://rm.coe.int/cai-2023-18-consolidated-working-draft-framework-convention/1680abde66.

258. Centre for Data Ethics and Innovation, *Review into bias in algorithmic decision-making* (London: UK Government, Dept. for Science, Innovation and Technology, 27 November 2020), online: https://www.gov.uk/government/publications/cdei-publishes-review-into-bias-in-algorithmic-decision-making/main-report-cdei-review-into-bias-in-algorithmic-decision-making, section 2.3.

population that is sampled.[259] The other is popular or non-technical: skewing of output in a way that would be commonly perceived as inconsistent with values of fairness that are widely held in society.[260] This is the sense in which "bias" is predominantly used in this report.

There are numerous ways in which bias can enter automated decision-making processes. It may be present in the design of the algorithm on which the system is based.[261] It may be present in the data used to train or test a system.[262] It may be introduced by input data collected and processed during the operational phase once the system has been deployed that is unrepresentative or improperly derived.[263] It can also result from a failure of human oversight through assumptions based on conscious or unconscious biases.[264] As mentioned in an earlier chapter, feedback loops can result from training new versions of a system using biased data generated by a previous version as inputs. Feedback loops replicate and may amplify the effect of the bias in the original output.[265]

Bias may lead easily to discrimination, by which we mean adverse treatment that lacks objective justification. Many examples exist of artificial intelligence systems generating discriminatory results. One of the most notorious involved a system called Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), a tool developed to predict recidivism for the purpose of granting bail or probation and which was used for a time by some U.S. courts and law enforcement authorities.

While race was not an input variable for risk assessment, COMPAS disproportionately predicted a higher risk of re-offending for Afro-American accused and was incorrect in its predictions at approximately the same rate but in opposite ways for

259. *Ibid.* A systematic error is one that is not determined by chance and is introduced by inaccuracy of observation or measurement inherent in a system: Merriam-Webster Dictionary online: https://www.merriam-webster.com/dictionary/systematic%20error.

260. *Supra*, note 258.

261. *Ibid.*

262. *Ibid.*

263. *Ibid.*

264. *Ibid.*

265. *Ibid.*

Afro-American and Caucasian accused.[266] It predicted 23.5% of Caucasian accused who did not re-offend as higher risk, and 44.9% of Afro-American accused.[267] It labelled 47.7% of Caucasian accused who did re-offend as lower risk, and 28.0% of Afro-American accused who did so at lower risk.[268]

An algorithm used in Arkansas to allot home care time to persons receiving disability benefits cut the care aide hours under a state program for a large number of recipients without any change having taken place in their circumstances.[269] The algorithm analyzed approximately 60 factors. The developer considered it a more "rational" system than subjective assessment by humans. A very small difference in a numerical score for some factors, however, could make a large difference in the allotment of care hours for persons in need of home care.

Another notorious example of algorithmic discrimination concerned a system tested by Amazon to vet candidates for employment. The data used to train the system reflected applications submitted to Amazon over the previous 10 years. The majority of applicants had been male. This evidently resulted in the system learning to classify female gender as a negative trait and reject applications by women. The discovery of this bias and other problems with the data led Amazon to abandon the system.[270]

Discriminatory effects from artificial intelligence outputs may be much more subtle than in the COMPAS and the Amazon examples. They will often not relate directly to the legally prohibited grounds of discrimination. Yet they will often present discrimination that is socially unacceptable.

Ostensibly neutral input variables like postal codes, income, and educational level may be correlated with demographic patterns and become proxies for race and ethnicity, leading to outputs that impose disadvantages on racialized and minority

266. Julia Angwin, Surya Mattu and Lauren Kirchner, "There's software used across the country to predict future criminals. And it's biased against blacks." *ProPublica* (23 may 2016), online: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

267. *Ibid.*

268. *Ibid.*

269. Colin Lecher, "What happens when an algorithm cuts your health care," *The Verge* (21 March 2018), online: https://www.theverge.com/2018/3/21/17144260/healthcare-medicaid-algorithm-arkansas-cerebral-palsy.

270. Jeffrey Dastin, "Amazon scraps secret AI recruiting tool that showed bias against women" Reuters, 10 October 2018, online: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

populations.[271] For example, an algorithm used to determine where Amazon Prime same-day delivery service could be economically provided in the U.S was reportedly found to exclude urban areas with a high proportion of Afro-American residents. Race was not an input variable, of course. One of the input variables evidently was the proximity of addresses to product warehouses, however.[272] As a result of prevailing patterns in urban settlement and housing, this served as a proxy for race.

Sometimes those who are adversely affected in a discriminatory manner will have recourse under existing anti-discrimination legal frameworks to have those effects corrected or to obtain compensation. In other cases they will not. Canadian anti-discrimination law only takes account of complaints based on a finite list of prohibited grounds in the relevant enactment. If the ground of differentiation is not included in that list, it is not recognized as discrimination. The equality rights section of the *Canadian Charter of Rights and Freedoms* (section 15) takes account of a few forms of discrimination that are not specifically listed, but a remedy under the Charter is only available against the Crown or a governmental entity.

This chapter concerns a gap that is arguably present in the law, namely the absence of a civil remedy for harm in the form of unjustifiable adverse treatment resulting from the application of artificial intelligence that is not capable of redress within the human rights framework. It raises the question whether the law of tort should fill that gap.

Some explanation of Canadian law relating to discrimination is necessary before going on.

271. Jacquelyn Burkell and Jane Bailey, "Unlawful Distinctions? Canadian Human Rights Law and Algorithmic Bias" (2016/2018) Can. Y.B. Human Rights 217 at 219; Bathaee, *supra*, note 63 at 920; *White House Blueprint for an Artificial Intelligence Bill of Rights*, *supra*, note 234 at 26; Rosel Kim and Kristen Thomasen, *Women's Legal Education & Action Fund (LEAF) Submission to The Standing Committee on Industry and Technology on Bill C-27* (11 September 2023), online: https://www.leaf.ca/wp-content/uploads/2023/09/2023-09-11-LEAF-Submission-re-AIDA-final.pdf at 9-10.

272. David Ingold and Spencer Soper, "Amazon Doesn't Consider the Race of Its Customers. Should It?" *Bloomberg*, 21 April 2016, online: https://www.bloomberg.com/graphics/2016-amazon-same-day/. Amazon's data analysis may not have involved artificial intelligence as such, but the example indicates how algorithmic discrimination by proxy can arise.

B. Overview of Canadian Anti-discrimination Law

1. A Legal Definition of Discrimination

The Supreme Court of Canada approved the following definition in *CN v. Canada (Canadian Human Rights Commission)*, contained in a Royal Commission report on employment equity:

> Discrimination … means practices or attitudes that have, whether by design or impact, the effect of limiting an individual's or a group's right to the opportunities generally available because of attributed rather than actual characteristics ….[273]

Some forms of discrimination are unconstitutional. Many of these would overlap with forms of discrimination that are prohibited under federal, provincial, or territorial human rights legislation. An important difference is that constitutional remedies may only be sought against a governmental defendant or a public entity that is an emanation of government, while the human rights legislation applies to governmental and private conduct.

2. Unconstitutional discrimination under the equality rights section of the Charter

Section 15(1) of the *Canadian Charter of Rights and Freedoms* ("the Charter") declares that every individual has the right to equal protection and equal benefit of the law "without discrimination and, *in particular, without discrimination based on race, national or ethnic origin, colour, religion, sex, age or mental or physical disability*." The guarantee of equality in section 15(1) is qualified by section 1 of the Charter, which subjects the rights and freedoms in it to "such reasonable limits prescribed by law as can be demonstrably justified in a free and democratic society."

Legislation and governmental actions that are inconsistent with the constitutional guarantee of equality in section 15(1) and not saved by demonstrably justified "reasonable limits" under section 1 are invalid. The onus to demonstrate justification on the basis of reasonable limits is on the party defending the challenged law or action of a government or public authority.

While Canadian courts have wide powers to grant remedies based on the Charter, the Charter does not apply to the conduct of private individuals or private corporate entities, as stated earlier.

273. [1987] 1 S.C.R. 1114 at 1138-39. This quotation was actually part of a longer attempt to explain the term "systemic discrimination."

Section 15(1) ostensibly leaves the term "discrimination" open-ended, but it is judicially interpreted to recognize claims of discrimination based on the expressly listed grounds of race, national or ethnic origin, colour, religion, sex, age or mental or physical disability and what the courts have described as "analogous grounds." This is obviously a rational interpretation of the term "discrimination" to prevent any law from being overturned merely because it differentiates between classes of individuals in some manner.

As currently interpreted, the test of discrimination under section 15(1) of the Charter is whether:

(a) the challenged law or governmental action creates a distinction on its face or in its impact on the basis of a ground listed in s. 15(1) or an analogous ground; and

(b) the distinction imposes a burden or denies a benefit in a manner having the effect of reinforcing, perpetuating, or exacerbating disadvantage.[274]

Analogous grounds are considered to be those based on immutable personal characteristics or ones that are changeable only at unacceptable cost to personal identity, which the federal or a provincial government has no legitimate interest in expecting an individual to change in order to receive equal treatment under the law.[275] They are ones that are actually immutable or "constructively immutable," such as religion.[276]

The Supreme Court of Canada has recognized only four analogous grounds of unconstitutional discrimination besides those expressly listed in section 15(1): citizenship, marital status, sexual orientation, and off-reserve residence of Indigenous peoples.[277] Other courts have recognized adopted status, the manner of conception of a child, receipt of public assistance, and parental status as analogous grounds.[278]

274. *Québec (Attorney General) v. Alliance du personnel professional et technique de la santé et des services sociaux*, 2018 SCC 17 at paras 25-28; *Centrale des syndicats du Québec v. Québec* (Attorney General), 2018 SCC 18 at para. 22; *Fraser v. Canada*, 2020 SCC 28 at para. 27.

275. *Corbière v. Canada*, [1999] 2 S.C.R. 203 at para. 13.

276. *Ibid.*

277. Hogg, *Constitutional Law of Canada*, 5th ed. Supplemented (Toronto: Thomson Reuters, 2022) at 55-56.

278. Errol Mendes and Stéphane Beaulac, eds. *Canadian Charter of Rights and Freedoms*, 5th ed. (Markham: LexisNexis, 2013) at 987.

3. Discrimination under human rights legislation

Federal, provincial, and territorial human rights legislation provides another forum for anti-discrimination claims. This legislation, unlike section 15(1) of the Charter, is binding on public entities and private persons alike.

Rather than providing court-based remedies, the pattern of current Canadian human rights legislation is one in which complaints of discrimination consisting of an alleged breach of the legislation are decided by a quasi-judicial tribunal. There may be a human rights commission or directorate interposed between the complainant and the tribunal. The human rights commission or directorate has an investigative and conciliation or mediation role as well as a general mandate to promote human rights. In some jurisdictions, the commission decides which complaints that it cannot settle have enough merit to proceed to adjudication by the tribunal. Thus, a complainant may not have an automatic right of access to adjudication, depending on the jurisdiction where the complaint arose. British Columbia is one of the jurisdictions that allow direct access by complainants to a Human Rights Tribunal.[279]

In order to be unlawful under human rights legislation, the discrimination that is the subject of a complaint must be based on a ground expressly prohibited by the legislation. An ostensibly neutral provision or rule may be discriminatory if it disproportionately affects a class of persons on the basis of a prohibited ground.[280] Intention to discriminate is not required to constitute a violation of human rights legislation, if the effect of the challenged action is discriminatory.[281]

Under the *Canadian Human Rights Act*, which applies to the federal government, federal agencies and the federally regulated private sector (including chartered banks, interprovincial railways, airlines, and marine industries), the prohibited grounds of discrimination are *race, national or ethnic origin, colour, religion, age, sex, sexual orientation, gender identity or expression, marital status, family status, genetic characteristics, disability, and conviction for an offence for which a pardon has been granted or in respect of which a record suspension has been ordered*.[282]

279. A single Human Rights Commissioner is appointed in British Columbia as an officer of the Legislature. The commissioner is responsible for promoting and protecting human rights, but does not control the case flow to the Human Rights Tribunal.

280. *British Columbia (Public Service Commission) v. British Columbia Government Employees Union*, [1999] 3 S.C.R. 3.

281. *Ontario Human Rights Commission v. Simpsons-Sears*, [1985] 2 S.C.R. 536.

282. R.S.C. 1985, c. H-6, s. 3(1).

The main prohibited grounds of discrimination under the British Columbia *Human Rights Code*, which are fairly typical of those in other provincial and territorial human rights enactments, are *Indigenous identity, race, colour, ancestry, place of origin, religion, marital status, family status, physical or mental disability, sex, sexual orientation, gender identity or expression, or age of that person or that group or class of persons*.[283]

Other prohibited grounds relating to specific provisions under the British Columbia *Human Rights Code* are: *the lawful source of income of a person or class* (concerning discrimination in residential tenancies,)[284] and in relation to employment, *a criminal or summary conviction offence that is unrelated to the employment or to the intended employment*[285] or *to the membership or intended membership*[286] of the complainant in a trade union.

When they find a complaint justified, human rights tribunals are typically empowered to make orders to cease or correct discrimination, or to adopt a plan or other measure to prevent future discrimination. They may also award monetary compensation to the complainant up to a ceiling amount.

The human rights tribunals have no jurisdiction over discrimination that is not based on a listed prohibited ground in the enactment under which they function.

4. No general tort of discrimination

In *Seneca College v. Bhadauria*, ("*Bhadauria*") the Supreme Court of Canada held that there is no common law intentional tort of discrimination, nor does a breach of human rights legislation give rise to an actionable tort.[287] The Supreme Court has maintained this position since *Bhadauria* was decided in 1981.[288] British Columbia courts continue to hold on the strength of the Supreme Court decisions that the existence of the comprehensive statutory scheme under the provincial *Human Rights*

283. R.S.B.C. 1996, c. 210, ss. 7(1), 8(1), 9(1).

284. *Ibid.*, s. 10(1).

285. *Ibid.*, s. 12(1).

286. *Ibid.*, s. 14.

287. [1981] 2 S.C.R. 181.

288. *University of British Columbia v. Berg*, [1993] 2 S.C.R. 353; *Honda Canada Inc. v. Keays*, 2008 SCC 39, at paras. 65-67; *Nevson Resources Ltd. v. Araya*, 2020 S.C.R. 5, [2020] 1 S.C.R. 166, at para. 240.

Code forecloses the possibility of a free-standing tort of discrimination at common law.[289]

British Columbia enacted a statute called the *Civil Rights Protection Act*[290] ("CRPA") shortly after *Bhadauria* was decided. It makes any conduct or communication a tort, actionable without proof of damage, if it has as its purpose interference with the civil rights of a person or class by promoting (a) hatred or contempt of a person or class, or (b) the superiority or inferiority of a person or class, on the basis of colour, race, religion, ethnic origin or place of origin.[291] The CRPA allows for awards of damages, exemplary damages, and injunctions.[292]

The statutory tort created by the little-used CRPA provides a remedy only against *intentional* conduct aimed at discrimination based on a few of the same grounds covered by the provincial *Human Rights Code*.[293] It would have no bearing on discriminatory effects of artificial intelligence outputs when these effects are unintended.

C. Discrimination Produced by Artificial Intelligence

1. Algorithmic Discrimination Without Remedy

Artificial intelligence systems can replicate biases in their training and input data that may go undetected in outputs for a considerable time, whether or not this is due to any blameworthiness on the part of developers and operators. The systems detect patterns in data that may lead them to make predictions and recommendations that, although not reached malevolently, are based on factors that may clash with legal or social norms of equality and fairness.

Where outputs of artificial intelligence produce discrimination on grounds covered by human rights legislation or section 15(1) of the *Charter*, those enactments may

289. *Olenga v. Royal Columbia Hospital*, 2017 BCSC 975 at para 12, aff'd 2018 BCCA 349; *Gichuru v. Law Society of British Columbia*, 2014 BCCA 396 at para, 103; *Schultz v. Beacon Roofing Supply Canada Company*, 2016 BCSC 1475.

290. R.S.B.C. 1996, c. 49. This Act seems not to have been passed specifically in response to *Bhadauria*, but to have been primarily aimed at Ku Klux Klan activity in British Columbia that took place shortly before its enactment: *Maughan v. UBC*, 2008 BCSC 14 at para. 333; aff'd 2009 BCCA 447; leave to appeal to S.C.C. refused 33495 (29 April 2010).

291. *Ibid.*, s. 2(1).

292. *Ibid.*, ss. 4(1), (3).

293. *Supra*, note 283.

provide a means of redress. If so, redress would need to be predicated on the proposition that the discriminatory behaviour of the systems could be attributed to their operators, developers, or someone who implements the discriminatory output. This is an unsettled area.[294]

Systems are often created to differentiate between human subjects, and to score or assess them on the basis of factors that are not prohibited grounds of discrimination under the Canadian human rights framework. Bias may affect these outputs, but the definition of "biased output" in the proposed Canadian *Artificial Intelligence and Data Act*[295] would limit the meaning of that term for the purposes of that Act to output that adversely differentiates individuals on grounds prohibited by the *Canadian Human Rights Act*.[296] This seems to preclude any remedy under the regulatory scheme of the *Artificial Intelligence and Data Act* if the adverse differentiation is not based on a prohibited ground under the federal human rights statute.

A system might be built for a lending institution to predict loan default risk for loan applicants, basing its recommendation to refuse a mortgage loan on analysis of large-scale datasets taking into account factors like the applicant's current residence in a particular district, because the loan loss ratio on a district-wide basis is one of the variables in the algorithm. It might weigh this variable more heavily than the applicant's own impeccable credit history. If the loan applicant is refused on basis of an address, the applicant has arguably suffered unfairly discriminatory treatment in relation to opportunities for housing under the definition of "discrimination" adopted by the Supreme Court of Canada in *CN v. Canada (Canadian Human Rights Commission)*, yet would appear to have no human rights remedy.

Two actual cases from the Netherlands and the UK illustrate discriminatory effects produced by automated decision-making that would probably not attract a human rights-based remedy in a Canadian jurisdiction:

- The Dutch government's SyRI system designed to predict the likelihood of tax or benefit fraud cross-analyzed data from several government departments to predict an individual's likelihood to commit tax or benefit fraud. It was found to target poorer neighbourhoods without any evidence of individual wrongdoing. SyRI was shut down by the order of a Dutch court on the basis of

294. See Burkell and Bailey, *supra*, note 271.

295. *Supra*, note 1.

296. The definition of "biased output" was likely limited to grounds of discrimination recognized by the *Canadian Human Rights Act*, supra, note 282 for constitutional reasons, namely to avoid an appearance of invading provincial jurisdiction.

inadequate protection for privacy and possible discrimination on the basis of socioeconomic or migrant status, which are not prohibited grounds under Canadian human rights legislation.[297]

- A standardization algorithm briefly used in the U.K. when examinations were cancelled during the pandemic assigned A-level and GCSE grades to students based on the historic performance of individual secondary schools. The system was intended to combat anticipated grade inflation from teachers predicting the marks their own students would have achieved. In England, Wales, and Northern Ireland, 36 per cent of secondary school students received an assigned grade lower than their teacher-predicted ones, and 3 percent two grades lower.[298] In Scotland, 25 per cent of teacher-predicted grades were reduced.

  The algorithm was found to have a tendency to decrease the grades of students in state schools and increase those in private schools because of its performance in relation to smaller population groupings.[299] As a result, many students lost out in university placements. A country-wide uproar resulted in apologies from the Westminster and Scottish governments and abandonment of the standardization algorithm.[300]

As usage of artificial intelligence expands, there will be more cases of unintended discriminatory effects that do not fit readily into the human rights framework, and the gap in the law into which these cases fall will become more apparent. They will be seen nevertheless as unfair and untenable, calling for some means of legal redress.

297. Jo Henley and Robert Booth, "Welfare surveillance system violates human rights, Dutch court rules" (2020), *The Guardian*, 5 Feb. 2020, online: https://www.theguardian.com/technology/2020/feb/05/welfare-surveillance-system-violates-human-rights-dutch-court-rules. The court also criticized the system's "serious lack of transparency."

298. "Sean Coughlan, Katherine Sellgren, Judith Burns, "A-Levels: Anger over 'unfair' results this year" BCC News, 13 August 2020, online: https://www.bbc.com/news/education-53759832.

299. Georgina Lee, "Did England exam system favour private schools?" *Channel 4 News*, 17 August 2020, online: https://www.channel4.com/news/factcheck/factcheck-did-england-exam-system-favour-private-schools.

300. "A-levels and GCSEs: How did the exam algorithm work?" BBC News, 20 August 2020, online: https://www.bbc.com/news/explainers-53807730 .

2. How Should the Gap in Protection Against Algorithmic Discrimination Be Closed?

(a) Human rights vs. tort remedy

Expanding the human rights framework to accommodate claims of discrimination resulting from the operation of artificial intelligence seems unlikely to be a satisfactory solution. Despite a tendency of some Canadian legislatures to expand the prohibited grounds in human rights statutes, the concept of discrimination under human rights legislation is primarily circumscribed by reference to characteristics of identity that are either "immutable or changeable only at unacceptable cost to personal identity,"[301] namely ones such as race, place of origin, ethnicity, age, disability, gender, language, and religion. The kinds of discrimination that artificial intelligence may create will frequently be based on other factors.

Discrimination through artificial intelligence may occur through attribution of characteristics to individuals by the system on the basis of factors that are not immutable such as place of residence, income level, or habits of consumption, and classification of individuals according to the attributed characteristics, such as lack of creditworthiness, or unsuitability for employment. While fitting within the generalized definition of discrimination enunciated by the Supreme Court of Canada in *CN v. Canada (Canadian Human Rights Commission)*, these forms of discrimination do not fit readily into the existing remedial framework of human rights legislation unless the basis of differentiation can be proven to have served as a proxy in the result for a prohibited ground of discrimination such as race, gender, or age.[302] This will not always be the case, but even when a non-prohibited ground has become a proxy for a prohibited ground of discrimination, requiring a complainant to prove this as a precondition for a human rights remedy would be onerous. It could potentially impose a greater resource strain on human rights tribunals and commissions in addition.

These additional difficulties of proof, and the lack of a remedy for some forms of algorithmic discrimination altogether under both the human rights framework and existing tort law, raise an issue of access to justice that will take on increasing importance with the expansion of artificial intelligence. The question arises whether a

301. *Corbière v. Canada*, *supra*, note 275 per McLachlin and Bastarache JJ. at para. 13, writing for the majority in the Supreme Court. While said in relation to s. 15(1) of the *Charter*, this description of the grounds on which unequal treatment is proscribed under the Charter is also largely apt with reference to prohibited grounds of discrimination that circumscribe the scope of other Canadian human rights legislation.

302. See *supra*, note 271 regarding superficially neutral input variables potentially becoming proxies for various prohibited grounds of discrimination.

civil remedy in tort should be made available for discriminatory treatment resulting from the operation of artificial intelligence.

It is arguable that civil courts are better adapted to deal with the technical complexity and lack of transparency of artificial intelligence decision-making than human rights tribunals because of their more comprehensive pre-trial oral and documentary discovery process. The pre-hearing discovery process of human rights tribunals generally requires only an exchange of documents.

Moreover, algorithmic discrimination is likely to have impacts differing from ones typically dealt with by human rights tribunals. Human rights complaints primarily concern affronts to personal dignity and mental distress resulting from them. Algorithmic discrimination is more likely to result in harm of a different kind, such as economic harm from arbitrary refusal of credit to a certain class of persons, or possibly increased exposure to physical health risks because of arbitrary classifications making certain groups of patients ineligible for a public or private health benefit. These are closer to the kinds of claims that civil courts are accustomed to ruling upon and quantifying.

A tort of algorithmic discrimination resulting from failure to take reasonable care would encourage designers, developers, and operators of artificial intelligence systems to take reasonable steps to detect and strive to eliminate unwanted bias in algorithms and data, and to monitor outputs for bias amongst other inaccuracies. Recall that these measures were identified in Chapter 5 as being widely recognized elements of good practice in the field of artificial intelligence that courts should consider in relation to the standard of care in claims based on negligence. Some recognized techniques and tools exist for detection, measurement, and treatment of unwanted bias in automated systems throughout their lifecycle.[303] The point was made in a response to the consultation paper that currently existing tools for data cleaning and bias detection should not be assumed to be fully reliable. Nevertheless, upstream and downstream defendants should be expected to employ available techniques to prevent or at least reduce the scope for harm from biased outputs.

Common law can evolve incrementally at times to recognize new torts. For this to occur, several criteria must be met.[304] The defendant must have caused some form

303. E.g., ISO/IEC/TR24027:2021 *Information technology - Artificial intelligence (I) – Bias in AI systems and AI aided decision making* issued by the International Standards Organization (ISO).

304. *Nevsun Resources Ltd. v. Araya*, 2020 SCC5 ("*Nevsun*") at para. 237 per Brown and Rowe, JJ. dissenting in part in the result. *Nevsun* concerned the question of whether Canadian common law provides a civil remedy against a private party for violation of customary international law. The majority held that it was not plain and obvious that the plaintiffs' claims had no reasonable

of unjustified harm to the plaintiff.[305] The new tort must be necessary to address the harm because adequate alternative remedies are lacking.[306] It must not amount to a "radical shift in the law."[307] Courts generally consider their role in revising the common law to be limited to incremental change involving extension of an existing principle to new circumstances, leaving major revisions of the common law that have complex ramifications to the legislature.[308]

Canadian courts have recognized new torts on several occasions in recent years. In 2007, the Supreme Court of Canada held that law enforcement authorities could be civilly liable for negligent investigation if compensable harm resulted from their conduct.[309] The Ontario Court of Appeal declared a new common law privacy tort of "intrusion upon seclusion" in 2012.[310] Ontario courts have recently recognized a new intentional tort of online harassment.[311] Alberta courts have recognized new torts of harassment[312] and public disclosure of private facts.[313] In addressing these forms of harm rather than discrimination and human rights, however, the Ontario and Alberta courts were not faced with a Supreme Court of Canada decision like *Bhadauria*[314] standing in their way.

prospect of success and for this reason should not be struck out on a preliminary application, and did not address the criteria for incremental recognition of new non-statutory torts.

305. *The Queen (Canada) v. Saskatchewan Wheat Pool*, [1983] 1 S.C.R. 205 at 224-225; *Nevsun*, *supra*, note 304 at para. 237

306. *Non-Marine Underwriters, Lloyd's of London v. Scalera*, 2000 SCC 24, [2001] 1 S.C.R. 551 (existing common law tort of battery adequate to cover cases of sexual battery without need to recognize a new tort of sexual battery that would incidentally require plaintiff to prove lack of consent); *Frame v. Smith*, [1987] 2 S.C.R. 99 (tortious remedy for interference by custodial parent with non-custodial parent's access rights precluded by existence of a comprehensive statutory scheme governing child custody and access following marital breakdown and providing remedies for enforcement of order granting access). See also *Nevsun*, *supra*, note 304 at paras. 237-240.

307. *Wallace v. United Grain Growers Ltd.*, [1997] 3 S.C.R. 701, at paras. 76-77; see also *Nevsun*, *supra*, note 304 at paras. 237 and 242, per Brown and Rowe, JJ., dissenting in part in the result.

308. *Winnipeg Child and Family Services (Northwest Area) v. G. (D.F.)*, [1997] 3 S.C.R. 925 at para. 18; *Watkins v. Olafson*, [1989] 2 S.C.R. 750 at 760-761; *R. v. Salituro*, [1991] 3 S.C.R. 654 at 670.

309. *Hill v. Hamilton-Wentworth Regional Police Services Board*, 2007 SCC 41, [2007] 3 S.C.R. 129.

310. *Jones v. Tsige*, 2012 ONCA 32, at para. 65.

311. *Caplan v. Atas*, 2021 ONSC 670; *385277 Ontario Ltd. v. Gold*, 2021 ONSC 4717.

312. *Alberta Health Services v. Johnston*, 2023 ABKB 209.

313. *E.S. v. Shillington*, 2021 ABQB 739, at para. 68.

314. *Supra*, note 287.

Writers have advanced compelling arguments that the tort of negligence is adaptable to address discrimination resulting from the operation of algorithms, and that *Bhadauria* is not a barrier because it concerned intentional racial discrimination.[315] Nevertheless, the provincial superior and appellate courts have shown considerable deference to *Bhadauria* and little interest in recognizing a right to sue in tort for discrimination. Legislation *may* be needed to create a remedy in tort for algorithmic discrimination.[316]

The parameters of a remedy in tort for algorithmic discrimination must be carefully set either by the legislature or the courts. If they are not, discrimination could be alleged in relation to virtually any differential treatment whatsoever, opening the door to indeterminate liability. This is among the principal reasons why Canadian courts have so far rejected the concept of a free-standing tort of discrimination.

(b) The Project Committee's view

(i) Negligent algorithmic discrimination

As unintended algorithmic discrimination typically results from a failure to eliminate unwanted bias or lack of care in the collection and treatment of data, the Project Committee concluded that a civil remedy for it should be based on negligence, whether the remedy is enacted as a statutory tort or introduced by judicial decision as an incremental change in the common law.

We would consider the essence of the proposed cause of action as a failure to take reasonable steps to detect and correct biased output of an artificial intelligence system or other algorithmic process, resulting in discrimination against a person or class that is either: (a) illegal because it is based on a ground prohibited by the Charter or other laws, or, (b) not based on a prohibited ground but is not warranted by reasonable business or industry practice. In the case of algorithmic discrimination on the basis of a prohibited ground under the Charter or other laws, human rights and Charter remedies might overlap with the proposed statutory tort. The mere existence of overlapping remedies would not permit duplicative compensation, however. It is often the case that the same conduct or set of facts gives rise to

315. Khoo, *supra*, note 125 at 53-55. Khoo distinguishes *Bhadauria* on the ground that it concerned intentional racial discrimination. Ruparelia has also distinguished *Bhadauria* on this ground in arguing that the case does not preclude recognition of a general tort of negligent discrimination: Rakhi Ruparelia, "I Didn't mean it That Way!": Racial Discrimination as Negligence" (2009) 44 S.C.L.R. (2d) 81.

316. Thomasen, *supra*, note 163 at 121.

concurrent remedies in contract and tort, or more than one type of tort, but the plaintiff is not compensated twice for the same loss or damage.[317]

For a negligence claim to succeed, there must be proof of damage. Maintaining this requirement in connection with *unintended* discrimination that is outside human rights legislation or the *Charter* would serve to prevent open-ended liability for any differentiation affecting an individual or class that flows from the output of a system. In other words, there should be proof of adverse consequences for the plaintiff beyond the mere fact of differential treatment. This is inherent in the legal meaning of "discrimination," but it would be the adverse consequences for the plaintiff, rather than the mere fact of differentiation, that would give rise to the right to sue. Additionally, the relative severity of the adverse consequences could be taken into account in assessing damages.

It may be that some kinds of harm that algorithmic discrimination may produce do not fit easily within the range of compensable damage in tort because they may, for example, involve loss of opportunity. These may be speculative heads of damage and difficult to quantify. Given the innumerable ways in which artificial intelligence is already affecting individual lives and will do so to an ever-increasing degree, however, courts should be prepared to take an expansive view of what amounts to actual damage from discrimination by machine.[318]

317. See, for example, *Achor v. Ihekweme*, 2023 ABKB 606, at paras. 52 and 55. Online defamation of the plaintiffs was held in that case to amount also to a recently recognized tort of harassment in addition, but the damages for harassment were subsumed in those awarded for defamation and there was no separate award for harassment. Similarly, a claim by the plaintiff for intentional infliction of mental suffering did not result in a separate award of damages because the same conduct was involved in the defamation claim.

318. Distress resulting from the fact of discrimination can give rise to considerable psychological harm even without additional damage of other kinds. Serious and prolonged psychological harm that goes beyond transitory psychological upset is already recognized as a form of personal injury that may be compensable in tort, if harm of this kind was a reasonably foreseeable consequence of the defendant's conduct. In *Mustapha v. Culligan of Canada Ltd.*, *supra*, note 70, McLachlin, C.J.C. stated at para. 8:

> Generally, a plaintiff who suffers personal injury will be found to have suffered damage. Damage for purposes of this inquiry includes psychological injury. The distinction between physical and mental injury is elusive and arguably artificial in the context of tort.

See also *Saadati v. Moorehead*, [2017] 1 S.C.R. 543, at para. 35. If a plaintiff shows that serious and prolonged psychological harm has resulted from algorithmic discrimination or deliberate failure to address known bias, this could amount to actual damage. There would need to be more than mere psychological upset that does not result in impairment of cognitive function, interference with daily living, or treatment for emotional symptoms: *Bothwell v. London Health Sciences Centre*, 2023 ONCA 323.

A major legal organization responding to the consultation paper expressed approval of a negligence-based tort of algorithmic discrimination in principle, but cautioned that the existence of a duty of care owed to the plaintiff and the lack of any policy grounds for negating that duty, which are essential elements in other negligence claims, should remain essential elements for the new tort as well.

Recommendations 8 and 9 below presuppose that the elements of negligence besides the ones they describe would need to be present for a non-statutory claim of algorithmic discrimination to succeed.[319] Regarding the requirement of a duty of care in the case of upstream defendants, we would point to the analogy made in Chapter 3 to the duty of care of manufacturers of a potentially hazardous, multi-component product, which is owed to the entire class of persons who reasonably may be foreseen to be affected after the product is released into the market.[320]

In the case of downstream defendants, we anticipate that the use of the system to make classifications and distinctions that can foreseeably affect individuals and classes adversely would create the degree of proximity necessary to give rise to a duty of care. The boundaries of the foreseeable class of those potentially affected will be determined by the circumstances. They may be wide or narrow, depending on the facts of a given use case, but again the constantly expanding influence of artificial intelligence into nearly every facet of life should lead courts to employ a careful but decidedly remedial approach in identifying them.

(ii) Intentional algorithmic discrimination

Society has a strong interest in discouraging the intentional design or use of artificial intelligence to produce discrimination. There is an equally strong societal interest in encouraging efforts to remove or mitigate the effects of bias that is capable of producing discrimination.

When the algorithmic discrimination was actually intended by defendants,[321] or if the defendants were aware of the bias and deliberately neglected to take reasonable steps to correct it, the deterrent function of tort would be served by introducing a presumption of actual damage in these circumstances, so that the fact of

319. If the tort of algorithmic discrimination were to be introduced by statute rather than by incremental development of common law, then of course the elements of the tort would be determined by the governing legislation.

320. *Bow Valley (Bermuda) Ltd. v. Saint John Shipbuilding Ltd.*, *supra*, note 142.

321. Deliberate creation of deepfaked photographs has recently become a prominent example of the intentional use of artificial intelligence for potentially discriminatory purposes.

discrimination from biased output would in itself give rise to a right to sue.[322] The Project Committee is prepared to recommend accordingly.[323]

(c) Recommendations

The Project Committee recommends:

8. (1) Failure to take reasonable steps to detect and correct biased output of an artificial intelligence system, resulting in discrimination against a person or class that is not warranted by reasonable business or industry practice, or is on a ground prohibited by law, should amount to actionable negligence if the differential treatment of the plaintiff resulting from the output is accompanied by actual damage.

(2) Recommendation 8(1) does not contemplate duplicative compensation if a plaintiff suffers discrimination on a ground prohibited by law resulting from biased output of an artificial intelligence system and has additional remedies in respect of the discrimination.

9. If the defendant intended the result described in Recommendation 8(1), or was aware of the biased output and deliberately failed to take reasonable steps to correct it, the differential treatment of the plaintiff should be conclusively presumed to have caused actual damage.

322. An objection to a presumption of actual damage was made in one response to the consultation paper on the ground that such a presumption is not in keeping with general tort principles. Actual damage is presumed in certain torts such as trespass and defamation in the form of libel, however. Furthermore, a presumption of actual damage does not imply substantial damages will be awarded. Recommendation 9 does not preclude an award of nominal damages in an appropriate case.

323. A plaintiff may need to weigh the advantage of suing on the basis of intentional algorithmic discrimination and thereby avoiding the need to prove actual damage against reduced prospects of recovery if the defendant's liability insurance will not be available to satisfy a judgment. Liability insurance policies typically exclude damage intentionally inflicted by the insured from coverage. See *Butterfield v. Intact Insurance Company*, 2023 ONCA 246. For this reason, it might be more advantageous for plaintiffs to frame their pleadings in negligence even when the defendants appear to have acted intentionally.

Chapter 7. Conclusion

The law of tort evolved over the course of centuries in the context of harmful interactions between humans in society. Artificial intelligence has brought about a very new context, one in which self-directing machines acting autonomously may cause harm to humans, to their property, or other interests protected by law.

This report sets out a series of recommendations for adapting the law of tort to deal with this new context. They were reached after a great deal of deliberation and full consideration of excellent submissions received in response to a detailed consultation paper. The recommendations are designed to be implemented either by courts employing the processes by which common law has always developed incrementally to deal with changing conditions and new situations, or by legislatures if policymakers and legislators see fit to enact them to expedite change in the law. Regardless of the mode of implementation, BCLI and the Project Committee are confident the recommendations present a necessary and balanced prescription for the evolution of the law of torts in an increasingly automated world.

Appendix

List of Recommendations

1. Civil liability for harm caused by artificial intelligence should not be based on strict liability. (p. 46)

2. Product liability principles should be adapted by analogy to determine rights and liabilities as between a plaintiff harmed by the operation of an artificial intelligence system and defendants who participated in the development of the system and in making it available for use, by treating

(a) the plaintiff similarly to a plaintiff claiming to have incurred loss or damage from a product comprising multiple components;

(b) developers of the system as owing a duty of care similar to that owed by a manufacturer of a complex product involving multiple integrated components towards persons or entities who foreseeably could be affected by a defect making the product dangerous;

(c) developers of components of the system as owing a duty of care similar to that owed by a supplier of a component of a complex product towards persons who foreseeably could be affected by a defect in the component that makes the component and the product in which it is integrated dangerous. (pp. 46-47)

3. (1) An individual or corporate entity with decision-making authority of a managerial nature over the operation of an artificial intelligence system and who thereby is in a position to exert some degree of control over the risk associated with its operation should be treated as an operator for the purpose of civil liability.

(2) A person or corporate entity described in paragraph (1) does not cease to be an operator merely because the operation of the artificial intelligence system in question is overseen or controlled by another artificial intelligence system. (p. 50)

4. The liability of operators and other persons who provide services in connection with the operation of an artificial intelligence system should be based on general principles of the law of negligence, subject to the recommendations made below. (p. 50)

5. Except as against any defendant who is found to have exercised reasonable care in the circumstances leading to an action for damages or other relief due to harm to persons or property arising from the operation of artificial intelligence, a court deciding such an action should be justified in drawing an inference that a lack of reasonable care on the part of defendants responsible for the design, development, training, testing, or use of the system is causally linked to the harm incurred by the plaintiff, if

(a) the harm alleged by the plaintiff is proven to have been caused by the output of the artificial intelligence system, either functioning alone or as a component of an integrated system;

(b) the evidence taken as a whole does not establish the exercise of reasonable care by defendants in the design, development, training, testing, and use of that system or yield an explanation for the behaviour of the system in the circumstances of the case that is consistent with the exercise of reasonable care by those defendants; and

(c) due to the characteristics of the artificial intelligence system, the plaintiff cannot reasonably be expected to identify specific acts or omissions by specific defendants that caused or materially contributed to causing the system to occasion the harm. (p. 66)

6. Where harm results from the operation of an artificial intelligence system and a claim based on negligence is made, the test of reasonable foreseeability of harm should be applied with regard to the risk that the system might behave unpredictably to cause harm in an unknown manner, taking into account

(a) attributes of the system known at the relevant time;

(b) intended use of the system; and

(c) known or predictable alternate uses of the system. (p. 71)

7. (1) The standard of care in litigation concerning artificial intelligence should be set so as to encourage responsible innovation and development, encompassing reasonable care to avoid foreseeable injury or loss.

(2) In setting the standard of care, courts should employ a broad interjurisdictional perspective, extending where applicable to

(a) nationally and internationally recognized best practices; and

(b) national and international regulatory standards

in the design, development, and operation of artificial intelligence systems. (p. 87)

8. (1) Failure to take reasonable steps to detect and correct biased output of an artificial intelligence system, resulting in discrimination against a person or class that is not warranted by reasonable business or industry practice, or is on a ground prohibited by law, should amount to actionable negligence if the differential treatment of the plaintiff resulting from the output is accompanied by actual damage.

(2) Recommendation 8(1) does not contemplate duplicative compensation if a plaintiff suffers discrimination on a ground prohibited by law resulting from biased output of an artificial intelligence system and has additional remedies in respect of the discrimination. (p. 106)

9. If the defendant intended the result described in Recommendation 8(1), or was aware of the biased output and deliberately failed to take reasonable steps to correct it, the differential treatment of the plaintiff should be conclusively presumed to have caused actual damage. (p. 106)

PRINCIPAL FUNDERS IN 2023

The British Columbia Law Institute expresses its thanks to its funders in 2023:

- Law Foundation of British Columbia
- Ministry of Attorney General
- Alzheimer Society of Canada
- BC Association of Community Response Networks
- The Council to Reduce Elder Abuse (CREA)
- Department of Justice Canada
- Notary Foundation
- Real Estate Foundation of British Columbia
- Simon Fraser University
- Vancouver Foundation
- McLachlin Fund

The Institute also reiterates its thanks to all those individuals and firms who have provided financial support for its present and past activities.

BRITISH COLUMBIA LAW INSTITUTE

Supported by

Ministry of Attorney General